

ÜBER DIE ANALYSE EINER KLASSE NICHTLINEARER SYSTEME

Dem Fachbereich Ingenieurwissenschaften
der Universität Erlangen-Nürnberg
zur Erlangung des Grades
Doktor der Ingenieurwissenschaften

vorgelegt von
Bernd Huber

Erlangen 1975

1975 № 378

ÜBER DIE ANALYSE EINER KLASSE NICHTLINEARER SYSTEME

Dem Fachbereich Ingenieurwissenschaften
der Universität Erlangen-Nürnberg
zur Erlangung des Grades
Doktor der Ingenieurwissenschaften

vorgelegt von
Bernd Huber

Erlangen 1975



Als Dissertation genehmigt von dem
Fachbereich Ingenieurwissenschaften
der Universität Erlangen-Nürnberg

Tag der Einreichung: 23.07.1975

Tag der Promotion: 5.11.1975

Dekan: Prof. Dr. rer. nat. H. J. Schneider

Berichterstatter: Prof. Dr.-Ing. R. Unbehauen

Dr. rer. nat. R. Böhme

INHALTSVERZEICHNIS

Einführung	7
I. MATHEMATISCHE GRUNDLAGEN	16
1. Räume	16
1.1 Metrische Räume	16
1.2 Vollständige metrische Räume	17
1.3 Lineare Räume	18
1.4 Normierte lineare Räume	19
1.5 Banachräume	20
1.6 Die $L_p^n(-\infty, \infty)$ -Räume	20
2. Operatoren	23
2.1 Die Verstärkung eines Operators	24
2.2 Die inkrementale Verstärkung eines Operators	26
2.3 Lineare Operatoren	27
2.3.1 Die Norm eines linearen, beschränkten Operators	27
2.4 Nichtlineare Operatoren	29
2.5 Die Inverse eines Operators	29
2.6 Operatoren und Systeme	31
2.6.1 Zeitinvariante Systeme	32
2.6.2 Gedächtnisfreie Systeme	33
2.6.3 Kausale Systeme	33
3. Fixpunkte von Operatoren	34
3.1 Fixpunkte	34
3.2 Kontraktionen	34
3.3 Der Banachsche Fixpunktsatz	35
4. Fourier-Transformation und Faltungsoperatoren	38
4.1 Fourier-Transformation von L_1 -Funktionen	40
4.2 Fourier-Transformation von L_2 -Funktionen	41
4.3 Lineare Operatoren $L_{2r}^n(-\infty, \infty) \rightarrow L_{2r}^n(-\infty, \infty)$	43
4.4 Lineare Operatoren vom Faltungstyp	47

II. THEORIE EINER KLASSE NICHTLINEARER SYSTEME	50
1. Einleitung	50
2. Bedingungen für die Lösbarkeit der Operatorgleichungen	52
2.1 Der Banachraum B^n	52
2.2 Die Operatorgleichungen	53
2.3 Eine Klasse $\Psi(\alpha, \beta)$ nichtlinearer Operatoren, für die Gl. (3) gilt	56
2.4 Allgemeine Bedingungen für die eindeutige Lösbarkeit der Operatorgleichungen	60
2.4.1 Lösung der Gleichung $e = f + H N f$	60
2.4.2 Lösung der Gleichung $e = N f + H f$	63
2.4.3 Minimierung der Kontraktionsschranken $r(C, T)$ und $q(C, T)$	64
2.4.3.1 Der Einfluß von α und β auf die Kontraktionsschranken	65
2.4.3.2 Minimierung von $r(C, T)$	66
2.4.3.3 Minimierung von $q(C, T)$	85
2.4.3.4 Diskussion und Vergleich mit den Ergebnissen von Sandberg	87
2.4.4 Stetigkeitsbetrachtungen	92
2.4.5 Allgemeines	94
2.5 Spezielle Bedingungen für die Lösbarkeit der Operatorgleichungen	97
2.5.1 Voraussetzungen und Vereinbarungen	97
2.5.2 Bedingungen für die Existenz der Operatoren $(I + H C_0)^{-1}$ und $(C_0 + H)^{-1}$	99
2.5.3 Lösung der Gleichungen $e = f + H N f$ und $e = N f + H f$	104
2.5.3.1 Die Operatorgleichungen bei passivem linearen und nichtlinearen Systemanteil	107
2.5.3.2 Die Operatorgleichung $e = f + H N f$ bei $n = 1$	114
2.5.3.3 Die Operatorgleichung $e = N f + H f$ bei $n = 1$	116
3. Allgemeine Systembetrachtungen	117
3.1 Komplexe Systemstrukturen	117
3.2 Angangsbedingungen	119
3.3 Systemapproximation	121
3.4 Philosophie der Gleichungstransformation	122

III. ANALYSE EINER KLASSE NICHTLINEARER ZEITVARIANTER ELEKTRISCHER NETZWERKE	125
1. Einleitung	125
2. Vereinbarungen	126
3. Einige Klassen nichtlinearer zeitvarianter elektrischer Netzwerke	126
3.1 Elektrische Netzwerke mit nichtlinearen zeitvarianten Widerständen	126
3.1.1 RLCÜF-Netzwerke mit passiven nichtlinearen zeitvarianten Widerständen	132
3.2 Elektrische Netzwerke mit nichtlinearen zeitvarianten gesteuerten Quellen	141
3.3 Elektrische Netzwerke mit nichtlinearen zeitvarianten Widerständen und gesteuerten Quellen	146
3.4 Gleichstromanalyse elektrischer Netzwerke	146
IV. NUMERISCHE ERGEBNISSE	150
1. Einleitung	150
2. Ein Analyseverfahren	151
2.1 Voraussetzungen und Problemstellung	151
2.2 Iteration in Bezug auf einen endlichen Zeitbereich	154
2.3 Realisierung des Verfahrens unter Verwendung eines Digitalrechners	160
2.3.1 Darstellung von Funktionen und Operatoren	160
2.3.2 Das Lösungsverfahren	165
2.3.3 Konvergenz und Genauigkeit	169
3. Analysebeispiele	173
3.1 Systeme mit konzentrierten Elementen	174
3.1.1 Beispiel K1	174
3.1.2 Beispiel K2	178
3.1.3 Beispiel K3	183
3.1.4 Beispiel K4	188
3.1.5 Beispiel K5	196
3.2 Systeme mit verteilten Elementen	204
3.2.1 Beispiel V1	204
3.2.2 Beispiel V2	208
3.3 Vergleich mit dem Verfahren von Bulirsch-Stoer	212

Zusammenfassung	219
Literaturverzeichnis	222
Lebenslauf	224

EINFÜHRUNG

Bei der Behandlung technisch-physikalischer Problemstellungen wird der Ingenieur häufig vor die Aufgabe gestellt, Vorgänge und Zusammenhänge in konkreten Anordnungen modellartig zu beschreiben, um dann anhand der Modelle tiefere Einsichten in die Eigenschaften und das Verhalten dieser Anordnungen gewinnen und quantitative Aussagen machen zu können. Dabei stellt das Modell eine Idealisierung der wirklichen Verhältnisse dar, die vorgenommen wird, um einerseits einen präzisen Ausgangspunkt für theoretische Untersuchungen zu schaffen und andererseits unwesentliche Nebeneffekte zu eliminieren. In vielen Fällen kann das Modell als ein System gemäß Abb. (1) verstanden werden, auf das gewisse Größen einwirken (Systemerregung x) und das am "Ausgang" gewisse Größen liefert (Systemreaktion y).

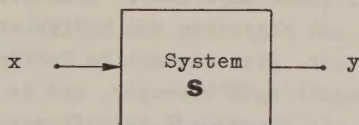


Abb. (1)

Dabei sind sowohl die Systemerregung als auch die Systemreaktion vektorwertige Funktionen der Zeit t , da das System im allgemeinen mehrere Ein- und Ausgänge besitzt. Das Gesamtsystem nach Abb. (1) ist häufig als Systemblockschaltbild gegeben, in dem verschiedene Teilsysteme zusammengeschaltet sind. Die Teilsysteme modellieren dabei untergeordnete Einflüsse bzw. Zusammenhänge, und das Verhalten des Gesamtsystems kann als Zusammenspiel dieser Teilsysteme gedeutet werden. Eine der Hauptaufgaben der Systemtheorie ist

es, bei der Ermittlung von Einsichten in und Aussagen über die Eigenschaften eines vorgegebenen Systems behilflich zu sein. Mit diesem Teil der Systemtheorie, der auch Systemanalyse genannt werden kann, beschäftigt sich diese Arbeit.

Gleichgültig in welcher mathematischen Form ein System gegeben ist (z.B. als Differentialgleichung), kann es als Zuordnung einer Funktion (Systemreaktion) zu einer vorgegebenen zulässigen Funktion (Systemerregung) verstanden werden. Diese Zuordnung ist natürlich im allgemeinen nicht nur für eine Erregungsfunktion, sondern für eine Menge Ω_1 von Erregungsfunktionen definiert. Ω_1 ist also die Menge der Funktionen, mit denen das System erregt werden kann, und wir nennen Ω_2 die Menge der Funktionen, welche dabei als Systemreaktion auftreten. Es ist sinnvoll anzunehmen, daß die Zuordnung eindeutig ist, da das System eine konkrete physikalische Anordnung beschreiben soll und der Ablauf eines reproduzierbaren physikalischen Experiments zumindest makroskopisch eindeutig ist und abgesehen von Meßfehlern eindeutige Meßergebnisse liefert. Eine eindeutige Zuordnung dieser Art heißt in der Mathematik auch Operator, und es liegt jetzt nahe, ein System als Operator S aufzufassen, der eine bestimmte Menge Ω_1 von Funktionen in eine Menge Ω_2 von Funktionen abbildet; man schreibt dafür

$$S: \Omega_1 \longrightarrow \Omega_2$$

und für die Systemreaktion y gilt symbolisch

$$y = Sx,$$

wobei $x \in \Omega_1$ die Systemerregung darstellt. Im folgenden werden die Begriffe System und Operator als Synonyme betrachtet. Der große Vorteil dieser sehr allgemeinen Auffassung von Systemen besteht nun darin, daß leistungsfähige Kon-

zepte aus der Mathematik wie, z.B. die Funktionalanalysis, für die Systemanalyse eingesetzt werden können.

Von besonderer Bedeutung ist in der Systemtheorie der Begriff der Linearität. Ein System heißt linear, wenn es die Anwendung des Superpositionsprinzips erlaubt [22]. In dieser Definition sind jedoch viele Systeme nicht enthalten, die der Ingenieur durchaus als linear ansehen würde (z.B. auf Grund der Anwesenheit "innerer" Systemerregungen oder nicht verschwindender Anfangszustände). Ein System dieser Art kann jedoch häufig durch ein lineares System ersetzt werden, zu dem am Ausgang eine Funktion addiert wird, die nur vom ursprünglichen System und nicht von der Systemerregung abhängt (siehe Abb. (2)).

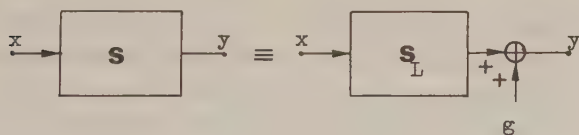


Abb. (2)

Die Eigenschaft der Linearität hat in der Systemtheorie deshalb so große Bedeutung, weil sich einerseits viele physikalisch-technische Gesetzmäßigkeiten durch lineare Systeme recht gut beschreiben lassen und andererseits die mathematische Behandlung solcher Systeme relativ einfach ist. Genau besehen weisen aber die meisten Gesetzmäßigkeiten die Linearitätseigenschaft nicht auf. Immer dann, wenn die Abweichungen von der Linearität gravierend sind, müssen wir mit einem nichtlinearen Modell arbeiten. Dabei verstehen wir unter nichtlinearen Systemen solche, die nicht notwendigerweise linear sind. (Lineare Operatoren sind also in

der Menge der nichtlinearen Operatoren enthalten.) Diese Wortschöpfung ist irreführend, sie ist aber üblich und wird deshalb beibehalten. Man weiß, daß die Fülle der Erscheinungen in nichtlinearen Systemen diejenigen Phänomene in linearen Systemen bei weitem übertrifft, und deshalb ist es nicht zu erwarten, daß es eine geschlossene Theorie gibt, mit der alle nichtlinearen Systeme behandelt werden können. Vielmehr wird man sich, wie es insbesondere in dieser Arbeit geschieht, auf eine Teilmenge nichtlinearer Systeme festlegen, und man erhält dann für den Preis gewisser Annahmen und Einschränkungen technisch interpretierbare Aussagen. Auf der anderen Seite ist es nicht sinnvoll, sich in den Voraussetzungen zu stark einzuschränken, da ja dann auch die Aussagen nur noch auf wenige Spezialfälle anwendbar sind.

In dieser Arbeit beschäftigen wir uns im wesentlichen mit einer Systemstruktur gemäß Abb. (3), die sich in der Praxis häufig dann ergibt, wenn das gesamte Systemverhalten durch das Zusammenspiel eines linearen und eines nichtlinearen Operators dargestellt werden kann (**H** und **N**).

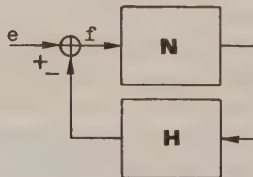


Abb. (3)

Dabei sei e die vorgegebene Systemerregung und f die gesuchte Systemreaktion. Die Verknüpfung zwischen e und f kann auch in Form der Operatorgleichung

$$e = f + \mathbf{H}\mathbf{N}f = (\mathbf{I} + \mathbf{H}\mathbf{N})f \quad (1)$$

angegeben werden. (Dabei ist $\mathbf{I}$ der identische Operator.) Da die Systemreaktion f in Gl. (1) implizit gegeben ist,

handelt es sich bei dem in Abb. (3) dargestellten System um ein Rückkopplungssystem. Damit wir möglichst universell anwendbare Aussagen treffen können, legen wir uns über die Art der Systemfunktionen e und f am Anfang unserer Untersuchungen nur insofern fest, als diese einem gewissen Banachraum angehören sollen, und nehmen an, daß die Operatoren H und N diesen Banachraum in sich abbilden. Für die technische Anwendung wählt man dann diesen Banachraum folgendermaßen:

- (B1) Raum der quadratisch integrierbaren Funktionen für die Untersuchung von Übergangsprozessen.
- (B2) Raum der periodischen und über eine Periode quadratisch integrierbaren Funktionen für die Untersuchung von periodischen Vorgängen.
- (B3) Raum der gewöhnlichen Vektoren für die Untersuchung von zeitlich konstanten Vorgängen (Analyse des eingeschwungenen Zustands).

Wir kommen jetzt zu zwei fundamentalen Fragestellungen, die sich wie ein roter Faden durch die ganze Arbeit ziehen:

- (F1) Unter welchen Bedingungen existiert das System gemäß Abb. (3) oder anders ausgedrückt: Unter welchen Bedingungen existiert für jede zulässige Systemerregung e eine eindeutige Systemreaktion f ?
- (F2) Wie läßt sich die Systemreaktion f berechnen, wenn das System existiert und die zulässige Systemerregung e gegeben ist?

Es sei betont, daß die Frage (F1) durchaus keinen "akademischen" Charakter hat, da, wie bereits weiter oben angedeutet wurde, an der Güte des Modells gezweifelt werden muß, wenn Gl. (1) keine oder gar mehrere Lösungen aufweist. Daß die Frage (F2) von technischem Interesse ist, braucht wohl nicht

näher betont zu werden. Beide Fragen lassen sich in die Frage nach der Existenz und Konstruktion des Systemoperators

$$\mathbf{S} = (\mathbf{I} + \mathbf{H}\mathbf{N})^{-1}$$

zusammenfassen. Diese und ähnlich geartete Problemstellungen wurden bereits unter gewissen Voraussetzungen zuerst von Zames [25] und später von Sandberg [9,10,11,12,13,14,15] untersucht. Beide lösten das gestellte Problem, indem sie die Systemgleichung in eine äquivalente Gleichung transformierten und auf diese den Banachschen Fixpunktsatz anwandten. Dieser Satz garantiert die eindeutige Lösbarkeit einer Operatorgleichung und erlaubt die Anwendung eines Iterationsverfahrens zur Berechnung der Lösung, wenn eine gewisse, aus der Operatorgleichung abgeleitete Zahl kleiner als 1 ist. Dabei ist die Konvergenzgeschwindigkeit des Iterationsverfahrens um so größer, je kleiner diese Zahl ist, die man auch Kontraktionskonstante nennt. Die genannten Transformationen erfolgten in der Weise, daß die der transformierten Operatorgleichung zugeordnete Kontraktionskonstante bezüglich einer ins Auge gefaßten Klasse von Transformationen minimal war [25,15]. Insbesondere betrachtete Sandberg Operatoren $\mathbf{N}$, die durch eine endliche Zahl nichtlinearer gewöhnlicher Funktionen definiert sind und zur Beschreibung von Systemen mit nichtlinearen Elementen dieser Art herangezogen werden können. Diese Ergebnisse wandte er dann auf die Analyse von elektrischen Netzwerken mit nichtlinearen zeitvarianten Widerständen an [12,14].

Nun zur vorliegenden Arbeit: In Kapitel I. werden mathematische Grundlagen behandelt, die einerseits für das Verständnis der folgenden Kapitel notwendig sind und andererseits im allgemeinen nicht zum mathematischen Rüstzeug des Ingenieurs gehören. Dabei handelt es sich um relativ einfache und bekannte Konzepte aus der Topologie und Funktional-

analysis. In Kapitel II. werden zunächst zwei Erweiterungen zu den Sandbergschen Arbeiten (insbesondere [15]) vorgenommen:

- (E1) Der nichtlineare Operator **N** besitzt Eigenschaften, die allgemeiner gefaßt und in der Praxis häufiger gegeben sind als die in den Sandbergschen Arbeiten vorausgesetzten.
- (E2) Die Klasse der möglichen Transformationen ist allgemeiner.

Mit diesen erweiterten Voraussetzungen, die als Spezialfall die Sandbergschen Voraussetzungen enthalten, wird eine Transformation vorgeschlagen, die im Hinblick auf eine minimale Kontraktionsschranke bezüglich einer sehr allgemeinen Klasse von Transformationen optimal ist. Diese Optimalitätsaussagen bilden einen wesentlichen Teil dieser Arbeit und können als direkte Verallgemeinerung der Sandbergschen Ergebnisse verstanden werden. Danach wenden wir diese Theorie durch eine geeignete Spezialisierung auf die Analyse von Übergangsvorgängen an, wobei wir als Banachraum den Raum der quadratisch integrierbaren Funktionen wählen. Von großer Bedeutung für die Anwendung auf die Netzwerktheorie ist dabei die Verallgemeinerung der Sandbergschen Arbeiten [12,13,14,15] auf den Fall, daß der lineare Operator **H** durch einen Faltungsoperator dargestellt wird, dessen Impulsantwort im Zeitnullpunkt einen Deltastoß enthalten darf. Der nichtlineare Operator **N** wird jetzt durch eine endliche Zahl nichtlinearer zeitvarianter Funktionen dargestellt, die gewisse Eigenschaften aufweisen. (Dadurch können z.B. bei der Anwendung auf die Netzwerktheorie nichtlineare zeitvariante Widerstände und gesteuerte Quellen dargestellt werden.) Unter diesen Voraussetzungen werden die allgemein formulierten Bedingungen für die Anwendbarkeit unserer Theorie in interessante technisch interpretierbare Kriterien umgesetzt. Diese Kriterien bezie-

hen sich im wesentlichen auf die Fourier-Transformierte der Impulsantwort des linearen Systemanteils und sind damit der Gedankenwelt des Ingenieurs angepaßt. Insbesondere sind die Kriterien grundsätzlich erfüllt, wenn die Systeme **H** und **N** passiv sind. Im Fall einer Nichtlinearität entsteht ein Frequenzgangkriterium, das an das Kreiskriterium der nichtlinearen Stabilitätstheorie erinnert, und es soll an dieser Stelle angedeutet werden, daß alle neuen Ansätze dieser Arbeit auch auf die Stabilitätsuntersuchungen nichtlinearer Systeme übertragen werden können [16,17,18,19,20]. In Kapitel III. werden die theoretischen Ergebnisse von Kapitel II. auf die Analyse einer Klasse elektrischer Netzwerke angewandt. Dabei betrachten wir Netzwerke, die neben linearen Elementen eine endliche Zahl nichtlinearer zeitvarianter Widerstände und gesteuerter Quellen enthalten, wobei die linearen Elemente nicht notwendig konzentriert sein müssen. Es wird gezeigt, daß sich diese Netzwerke durch Operatorgleichungen der Form Gl. (1) beschreiben lassen, wobei der Operator **H** den linearen Netzwerkanteil und **N** die nichtlinearen Widerstände und gesteuerte Quellen beschreibt. Dabei ist unsere Darstellung, über unsere bisherigen Verallgemeinerungen hinaus, allgemeiner als die Sandbergs [12,14], da sich dieser auf die Untersuchung von elektrischen Netzwerken mit nichtlinearen, zeitvarianten Widerständen beschränkt und dabei fordert, daß sich der lineare Netzwerkanteil durch eine Impedanzmatrix beschreiben läßt. Da wir von einer Hybridmatrixbeschreibung des linearen Netzwerkanteils ausgehen, die, wie wir wissen, für jedes RLCÜF-Netzwerk existiert [6], sind wir in der Lage zu beweisen, daß unter schwachen Bedingungen jedes Netzwerk, das neben nichtlinearen, zeitvarianten und passiven Widerständen aus RLCÜF-Netzwerkelementen besteht, für jede quadratisch integrierbare Netzwerkerregung eindeutige quadratisch integrierbare Ströme bzw. Spannungen in bzw. an den

nichtlinearen Widerständen existieren, welche die das Netzwerk beschreibende Gleichung erfüllen und anhand eines Iterationsverfahrens berechnet werden können. Die genannten schwachen Bedingungen beziehen sich dabei auf die Netzwerk-erregungen und sie sind in den meisten praktisch interessanten Fällen erfüllt. Den Abschluß von Kapitel III. bildet die Gleichstromanalyse elektrischer Netzwerke, die grundsätzlich mit unserer Theorie durchgeführt werden kann, wenn man den Banachraum geeignet wählt.

An die mehr mathematisch orientierten Untersuchungen in den Kapiteln II. und III. schließt sich in Kapitel IV. die Frage nach der günstigen Ermittlung numerischer Resultate an, wobei wir uns auf die Analyse von Übergangsvorgängen beschränken. Zunächst wird gezeigt, daß unter gewissen sinnvollen und technisch interpretierbaren Bedingungen das Iterationsverfahren in der Weise modifiziert werden kann, daß alle relevanten Funktionen nur in einem endlichen Zeitintervall berechnet werden, was vor allem im Hinblick auf die Darstellbarkeit von Funktionen auf dem Digitalrechner von Bedeutung ist. Danach wird das Iterationsverfahren in einen numerischen Algorithmus umgesetzt, mit dem verschiedene praktische Beispiele untersucht werden. Ferner unterziehen wir diesen Algorithmus einer Fehleranalyse anhand derer eine Abschätzung der unvermeidlichen Rechenfehler vorgenommen werden kann. Bei all diesen Beispielen zeigt es sich, daß die hinreichenden Bedingungen für die eindeutige Existenz der Lösung und die Konvergenz des Iterationsverfahrens keine "akademischen" Forderungen darstellen, sondern in vielen technisch interessanten Fällen erfüllt sind. Wenn die Kriterien nicht erfüllt sind, werden auch häufig die Aussagen unserer Theorie verletzt und in diesem Sinne sind die hinreichenden Bedingungen unter denen unsere Sätze gelten "fast notwendig". Im Anschluß an die Analysebeispiele wird unser Iterationsverfahren für den Fall, daß das lineare Teilsystem H aus konzentrierten Elementen besteht, mit einem bekannten Integrationsverfahren verglichen.

I. MATHEMATISCHE GRUNDLAGEN

1. Räume

Ein Raum ist eine Menge von Elementen (auch Punkte genannt), welche eine gewisse Struktur besitzt. Eine Struktur ist durch bestimmte Bedingungen an die Elemente und Relationen zwischen den Elementen gegeben. Erst die Struktur eines Raumes ermöglicht es, mit ihm Mathematik zu treiben. Je "mehr" Struktur man einem Raum aufprägt, desto weniger allgemein wird er und desto spezieller werden auch Aussagen und Sätze, die sich auf diesen Raum beziehen.

Da es unser Ziel sein wird, technisch interpretierbare Ergebnisse zu ermitteln, werden im folgenden einige Räume definiert, die in der Systemtheorie von Bedeutung sind. Insbesondere werden Räume betrachtet, deren Elemente Funktionen sind. Diese Räume nennt man Funktionsräume, und die Mathematik, die sich auf diese Räume bezieht, heißt Funktionalanalysis.

1.1 Metrische Räume

Ein metrischer Raum (X, d) ist eine nichtleere Menge von Punkten, in der eine Distanzfunktion d (auch Metrik genannt) definiert ist. Diese Funktion ordnet jedem Punktepaar $x, y \in X$ einen reellen nichtnegativen Wert $d(x, y)$ zu und muß folgende Eigenschaften haben (für alle $x, y, z \in X$):

- (i) $d(x, y) = 0 \iff x = y$,
- (ii) $d(x, y) = d(y, x)$ (Symmetrie),
- (iii) $d(x, z) \leq d(x, y) + d(y, z)$ (Dreiecksungleichung).

In einem metrischen Raum kann die Konvergenz einer Punktfolge definiert werden:

Def. 1: (X, d) sei ein metrischer Raum. Eine Folge $\{x_n\}$ mit $x_n \in X$ konvergiert gegen einen Grenzwert $\hat{x} \in X$, falls für jedes $\epsilon > 0$ ein $N(\epsilon)$ existiert, derart, daß $d(\hat{x}, x_n) < \epsilon$ für alle $n > N(\epsilon)$ gilt. Man schreibt dann

$$\hat{x} = \lim_{n \rightarrow \infty} x_n \quad \text{oder} \quad x_n \rightarrow \hat{x}.$$

Eine konvergente Folge kann nur einen Grenzwert haben:

Aus $x_n \rightarrow \hat{x}_1$ und $x_n \rightarrow \hat{x}_2$ folgt

$$d(\hat{x}_1, \hat{x}_2) \leq d(x_n, \hat{x}_1) + d(x_n, \hat{x}_2) \rightarrow 0.$$

Damit ist $d(\hat{x}_1, \hat{x}_2) = 0$, und per Definition der Metrik gilt $\hat{x}_1 = \hat{x}_2$.

1.2 Vollständige metrische Räume

Um bei einer vorgegebenen Folge $\{x_n\}$ mit $x_n \in X$ unter direkter Anwendung der Konvergenzdefinition die Frage nach der Konvergenz zu entscheiden, benötigt man bereits den Grenzwert (falls er existiert). Diese Schwierigkeit kann bei einer bestimmten Klasse metrischer Räume umgangen werden. Zunächst definieren wir:

Def. 2: Gegeben sei ein metrischer Raum (X, d) . Eine Cauchy-Folge ist eine Folge $\{x_n\}$, $x_n \in X$ mit folgender Eigenschaft: Für alle $\epsilon > 0$ gibt es ein $N(\epsilon)$ derart, daß für alle $n, m > N(\epsilon)$ die Ungleichung $d(x_n, x_m) < \epsilon$ gilt. Wir schreiben für diese Art der Konvergenz

$$\lim_{n, m \rightarrow \infty} d(x_n, x_m) = 0$$

Man sagt, eine Cauchy-Folge erfüllt das Cauchysche Konvergenzkriterium. Eine Cauchy-Folge ist also eine Folge, bei der die Folgenglieder gegeneinander konvergieren. Es ist aber keineswegs immer sichergestellt, daß eine Cauchy-Folge gegen ein Element $\hat{x} \in X$ konvergiert. Es existieren jedoch metrische Räume, in denen jede Cauchy-Folge gegen ein Element des Raumes konvergiert:

Def. 3: Ein metrischer Raum (X, d) heißt vollständig, wenn jede Cauchy-Folge einen Grenzwert in X hat.

Falls eine Folge konvergiert, ist sie auch eine Cauchy-Folge, da $d(x_n, x_m) \leq d(x_n, \hat{x}) + d(x_m, \hat{x}) \longrightarrow 0$ gilt. Damit gilt

Satz 1: Notwendig und hinreichend für die Konvergenz einer Folge in einem vollständigen metrischen Raum ist, daß sie das Cauchysche Konvergenzkriterium erfüllt.

1.3 Lineare Räume

Ein linearer Raum X ist eine Menge von Elementen, für die eine Operation "Addition" und eine Operation "skalare Multiplikation" definiert ist. Die Operation "Addition" weist jedem Punktepaar $x, y \in X$ einen Punkt des Raumes zu, während die Operation "skalare Multiplikation" jedem Punkt einen anderen Punkt zuordnet. Dabei müssen die genannten Operationen die folgenden Bedingungen erfüllen:

- (1) Es gibt für alle $x, y \in X$ ein eindeutig definiertes $z = x + y$ (+ bedeutet eine Verknüpfung, die Addition genannt wird), so daß $z \in X$ ist, und es gilt:

- (i) $x + y = y + x$,
- (ii) $x + (y + w) = (x + y) + w$ für alle $w \in X$,
- (iii) es gibt ein Element $\theta \in X$ derart, daß $x + \theta = x$ ist (θ ist das Nullelement),
- (iv) für alle $x \in X$ gibt es ein Element $-x \in X$ derart, daß $x + (-x) = \theta$ gilt.

(2) Für alle $\lambda \in K$ (Körper der reellen oder komplexen Zahlen) und alle $x \in X$ gibt es ein Element $\lambda x \in X$ (Produkt von λ und x), derart, daß

- (v) $\lambda(\mu x) = (\lambda\mu)x$ für alle $\mu \in K$,
- (vi) $1x = x$,
- (vii) $(\lambda + \mu)x = \lambda x + \mu x$ für alle $\mu \in K$ und
- (viii) $\lambda(x + y) = \lambda x + \lambda y$ für alle $y \in X$

gilt. Je nachdem, ob K die Menge der reellen oder komplexen Zahlen ist, sprechen wir von einem reellen oder komplexen linearen Raum.

1.4 Normierte lineare Räume

Ein linearer Raum X heißt normiert, wenn jedem Element x eine nichtnegative reelle Funktion $\|x\|$ zugeordnet ist, die folgende Bedingungen erfüllt:

- (i) $\|x\| = 0 \iff x = \theta$,
- (ii) $\|\lambda x\| = |\lambda| \cdot \|x\|$ für alle $\lambda \in K$,
- (iii) $\|x + y\| \leq \|x\| + \|y\|$ für alle $y \in X$ (Dreiecksungleichung).

Ein normierter linearer Raum wird, wie man leicht verifiziert, immer zu einem metrischen Raum, wenn man $d(x,y) = \|x - y\|$ setzt. (Ein metrischer Raum kann nicht immer zu einem normierten Raum gemacht werden.)

1.5 Banachräume

Da jeder normierte lineare Raum auch ein metrischer Raum ist, kann er auch vollständig sein. Er hat dann die Eigenschaft, daß jede Cauchy-Folge (d.h. hier $\|x_n - x_m\| \rightarrow 0$ für $n, m \rightarrow \infty$) gegen ein Element des Raumes konvergiert. Wir definieren:

Def. 4: Ein vollständiger normierter linearer Raum heißt (reeller oder komplexer) Banachraum.

1.6 Die $L_p^n(-\infty, \infty)$ -Räume

Die Menge der vektorwertigen (mit n komplexwertigen Komponenten) und im lebesgueschen Sinne meßbaren Funktionen der reellen Variablen t wird mit $K^n(-\infty, \infty)$ bezeichnet, wenn der Definitionsbereich das zweiseitig unendliche Intervall $R = (-\infty, \infty)$ ist. Die i -te Komponente jedes Elementes $x \in K^n(-\infty, \infty)$ wird mit x_i bezeichnet. $x(t)$ ist der Wert der Funktion an der Stelle t , während x für die ganze Funktion steht. Mit $L_p^n(-\infty, \infty)$ bezeichnen wir die Teilmenge von $K^n(-\infty, \infty)$, deren Elemente die Bedingung

$$\int_{-\infty}^{\infty} \sum_{i=1}^n |x_i(t)|^p dt < \infty \quad \text{für } 1 \leq p < \infty$$

und

$$\sum_{i=1}^n \operatorname{ess\,sup}_{t \in \mathbf{R}} |x_i(t)| < \infty \quad \text{für } p = \infty$$

erfüllen. (ess sup ist das Supremum auf $\mathbf{R}/N$, wobei N eine beliebige Nullmenge ist.) Man kann zeigen [8,24], daß die $L_p^n(-\infty, \infty)$ -Räume komplexe lineare Räume sind und mit der Norm

$$\|x\| = \left\{ \int_{-\infty}^{\infty} \sum_{i=1}^n |x_i(t)|^p dt \right\}^{1/p} \quad \text{für } 1 \leq p < \infty$$

bzw.

$$\|x\| = \sum_{i=1}^n \operatorname{ess\,sup}_{t \in \mathbf{R}} |x_i(t)| \quad \text{für } p = \infty$$

werden sie ferner zu normierten Räumen. Es kann gezeigt werden, daß diese Räume auch vollständig sind [8], und damit gilt folgender Satz:

Satz 2: Die Räume $L_p^n(-\infty, \infty)$ sind mit obiger Normdefinition komplexe Banachräume.

Bemerkung 1:

Wenn man die n Komponenten der Funktionen als reellwertig voraussetzt, erhält man mit denselben Normdefinitionen die reellen Banachräume $L_{pr}^n(-\infty, \infty)$.

Bemerkung 2:

Im Vorhergehenden wurde der Definitionsbereich der Funktionen willkürlich gleich $\mathbf{R}$ gewählt. Bei entsprechender Ab-

wandlung der Normen können auch andere Definitionsbereiche verwendet werden. So haben z.B. die Elemente der Banachräume $\mathbb{L}_p^n[0, \infty)$ den Definitionsbereich $[0, \infty)$.

Bemerkung 3:

Für $n = 1$ setzen wir $\mathbb{L}_p = \mathbb{L}_p^1$ für alle $p \geq 1$.

Bemerkung 4:

Die bei den Normdefinitionen vorkommenden Integrale sind im Lebesgueschen Sinne zu verstehen. Diese Vereinbarung ist sehr wesentlich, da der Satz 2 nur unter dieser Voraussetzung gültig ist. Der Ingenieur wird sich bei der praktischen Berechnung mit Funktionen begnügen, für die der Riemannsche und Lebesguesche Integralbegriff gleichbedeutend ist.

Bemerkung 5:

Wie unmittelbar aus der Normdefinition der $\mathbb{L}_p^n$ -Räume ersichtlich, werden zwei Punkte $x_1, x_2 \in \mathbb{L}_p^n$ als gleich angesehen, wenn sie sich nur auf einer Menge vom Maße Null unterscheiden, d.h. wenn sie fast überall gleich sind.

Bemerkung 6:

Die $\mathbb{L}_2^n$ -Räume nennt man auch Hilbert-Räume, da es möglich ist, aus der $\mathbb{L}_2^n$ -Norm sowohl eine Metrik als auch ein inneres Produkt abzuleiten.

2. Operatoren

Operatoren werden als Verallgemeinerung des Funktionsbegriffes der klassischen Analysis eingeführt:

X und Y seien zwei beliebige nichtleere Mengen. Es sei eine Abbildung F gegeben, die jedem Punkt x einer Teilmenge $\Omega \subset X$ genau einen Punkt $y = Fx$ der Menge Y zuordnet. Diese Abbildung oder Transformation nennen wir einen Operator F , der eine Menge Ω in die Menge Y abbildet, und schreiben dafür $F: \Omega \rightarrow Y$. Dabei ist Ω der Definitionsbereich und $\{y | y = Fx, x \in \Omega\}$ der Bildbereich des Operators. Falls der Bildbereich gleich Y ist, bildet der Operator die Menge Ω auf Y ab. (Man sagt, die Abbildung ist surjektiv.) Die Menge Y kann z.B. die Menge der reellen oder komplexen Zahlen sein. Operatoren dieser Art nennt man auch Funktionale. Wenn wir voraussetzen, daß X und Y metrische Räume sind, können wir den üblichen Begriff der Stetigkeit von Funktionen auf Operatoren erweitern:

Def. 5: Ein Operator F , der eine Teilmenge Ω eines metrischen Raumes X in einen metrischen Raum Y abbildet, heißt stetig im Punkte $\hat{x} \in \Omega$, falls aus $x_n \rightarrow \hat{x}$ mit $x_n \in \Omega$ immer $Fx_n \rightarrow F\hat{x}$ folgt.

Aus dem folgenden Satz ergibt sich eine gleichwertige Definition der Stetigkeit [4]:

Satz 3: Ein Operator F , der eine Teilmenge Ω eines metrischen Raumes X in einen metrischen Raum Y abbildet, ist genau dann im Punkte $\hat{x} \in \Omega$ stetig, wenn es zu jedem $\epsilon > 0$ ein $\delta(\epsilon, \hat{x}) > 0$ gibt, derart, daß aus $d(x, \hat{x}) < \delta$ und $x \in \Omega$ immer $d(Fx, F\hat{x}) < \epsilon$ folgt. (Dabei ist zu beachten, daß sich $d(x, \hat{x})$ auf die Metrik im Raum X und $d(Fx, F\hat{x})$ auf die Metrik im Raum Y bezieht.)

Falls $\mathbf{F}$ für alle $\hat{x} \in \Omega$ stetig ist, heißt $\mathbf{F}$ auf Ω stetig, und wenn δ nicht von $\hat{x}$ abhängt, heißt $\mathbf{F}$ gleichmäßig stetig.

Man kann auch Abbildungen hintereinander schalten. So bedeutet der Operator $\mathbf{F}_2 \mathbf{F}_1$ die Abbildung des Definitionsbereichs von $\mathbf{F}_1$ in den Bildbereich von $\mathbf{F}_2$. Dabei muß gewährleistet sein, daß der Bildbereich von $\mathbf{F}_1$ im Definitionsbereich von $\mathbf{F}_2$ enthalten ist. Man nennt den Operator $\mathbf{F}_2 \mathbf{F}_1$ die Komposition von $\mathbf{F}_1$ und $\mathbf{F}_2$.

Bemerkung 1:

Mit den Operatoren $\mathbf{F}_1 : \Omega \rightarrow \mathbf{Y}$ und $\mathbf{F}_2 : \Omega \rightarrow \mathbf{Y}$ kann ein Summenoperator $\mathbf{F}_1 + \mathbf{F}_2$ durch $(\mathbf{F}_1 + \mathbf{F}_2)x = \mathbf{F}_1x + \mathbf{F}_2x$ für alle $x \in \Omega$ erklärt werden, falls $\mathbf{Y}$ ein linearer Raum ist. Dann ist es auch möglich, vom Produktoperator $\lambda \mathbf{F}$ mit beliebigem $\lambda \in \mathbf{K}$ zu sprechen: $(\lambda \mathbf{F})x = \lambda(\mathbf{F}x)$ für alle $x \in \Omega$. ($\mathbf{K}$ ist der Körper des linearen Raumes.)

Bemerkung 2:

Die Theorie der Operatoren, deren Definitions- und Bildbereich Funktionsräume sind, heißt Funktionalanalysis.

2.1 Die Verstärkung eines Operators

Es seien $\mathbf{X}$ und $\mathbf{Y}$ normierte lineare Räume, und es bilde der Operator $\mathbf{F}$ eine Teilmenge Ω von $\mathbf{X}$ in $\mathbf{Y}$ ab. $\mathbf{F}$ heißt beschränkt, wenn es ein $M < \infty$ gibt, so daß für alle $x \in \Omega$

$$\|\mathbf{F}x\| \leq M \cdot \|x\| \quad (1)$$

gilt. M heißt obere Schranke der Abbildung. Häufig stellt sich die Frage nach der besten Abschätzung, d.h. nach der

kleinsten oberen Schranke, für die Gl. (1) noch gilt. Auf Grund des folgenden Satzes existiert diese kleinste obere Schranke, sie ist gleich dem Supremum der Streckungsfaktoren, und wir nennen sie die Verstärkung $g(\mathbf{F})$ des Operators $\mathbf{F}$.

Satz 4: $\mathbf{X}$ und $\mathbf{Y}$ seien zwei normierte lineare Räume.

$\mathbf{F}$ sei ein beschränkter Operator von $\Omega \subset \mathbf{X}$ in $\mathbf{Y}$.

Dann existiert

$$g(\mathbf{F}) = \sup_{\substack{x \in \Omega \\ x \neq \theta}} \frac{\|\mathbf{F}x\|}{\|x\|},$$

und es gilt

$$g(\mathbf{F}) = \min \{ M \mid M \text{ ist obere Schranke von } \mathbf{F} \}$$

Beweis:

Wegen $\|\mathbf{F}x\| \leq M \cdot \|x\|$ für alle $x \in \Omega$ existiert $g(\mathbf{F})$. Für alle $x \in \Omega$ mit $x \neq \theta$ gilt

$$\frac{\|\mathbf{F}x\|}{\|x\|} \leq g(\mathbf{F}).$$

Da $\mathbf{F}$ als beschränkt vorausgesetzt wurde, gilt $\mathbf{F}\theta = \theta$ (falls $\theta \in \Omega$) und somit für alle $x \in \Omega$ $\|\mathbf{F}x\| \leq g(\mathbf{F}) \cdot \|x\|$. $g(\mathbf{F})$ ist also obere Schranke der Abbildung.

Nach Definition von $g(\mathbf{F})$ gibt es zu jedem $\epsilon > 0$ ein $x_\epsilon \neq \theta$ mit $x_\epsilon \in \Omega$, so daß

$$\frac{\|\mathbf{F}x_\epsilon\|}{\|x_\epsilon\|} > g(\mathbf{F}) - \epsilon$$

oder $\|\mathbf{F}x_\epsilon\| > (g(\mathbf{F}) - \epsilon) \cdot \|x_\epsilon\|$ gilt. Damit ist kein $g(\mathbf{F}) - \epsilon$ mit $\epsilon > 0$ obere Schranke von $\mathbf{F}$ und folglich ist $g(\mathbf{F})$ die kleinste obere Schranke.

q.e.d.

2.2 Die inkrementale Verstärkung eines Operators

X und **Y** seien normierte lineare Räume, und der Operator **F** bilde die Teilmenge Ω von **X** in **Y** ab. **F** ist inkremental beschränkt (man sagt auch **F** ist im Lipschitzschen Sinne stetig), falls es ein nur von **F** abhängiges $M < \infty$ gibt, so daß für alle $x, y \in \Omega$

$$\|F_x - F_y\| \leq M \cdot \|x - y\| \quad (2)$$

gilt. Auch hier taucht die Frage nach der kleinsten inkrementalen Schranke M auf, für die Gl. (2) noch gilt. Analoge Überlegungen zu denen von Satz (4) führen uns zur inkrementalen Verstärkung $g_0(\mathbf{F})$ der Operators **F**:

Satz 5: **X** und **Y** seien zwei normierte lineare Räume.

F sei ein inkremental beschränkter Operator von $\Omega \subseteq \mathbf{X}$ in **Y**. Dann existiert

$$g_0(\mathbf{F}) = \sup_{\substack{x, y \in \Omega \\ x \neq y}} \frac{\|F_x - F_y\|}{\|x - y\|},$$

und es gilt

$$g_0(\mathbf{F}) = \min \{ M \mid M \text{ ist inkrementale Schranke von } \mathbf{F} \}.$$

Beweis:

Analog zum Beweis von Satz 4.

Falls mit $\theta \in \Omega$ $F\theta = \theta$ gilt, folgt aus der inkrementalen Beschränktheit von **F** die normale Beschränktheit, und es gilt $g(\mathbf{F}) \leq g_0(\mathbf{F})$.

Folgender Satz wird für uns noch von Bedeutung sein:

Satz 6: $\mathbf{X}, \mathbf{Y}$ und $\mathbf{Z}$ seien normierte lineare Räume. Der Operator $\mathbf{F} = \mathbf{F}_2 \mathbf{F}_1$ sei die Hintereinanderschaltung der Operatoren $\mathbf{F}_1: \Omega_1 \rightarrow \mathbf{Y}$ und $\mathbf{F}_2: \Omega_2 \rightarrow \mathbf{Z}$ mit $\Omega_1 \subseteq \mathbf{X}$, $\Omega_2 \subseteq \mathbf{Y}$ und $\{y | y = \mathbf{F}_1 x, x \in \Omega_1\} \subseteq \Omega_2$. Ferner seien $\mathbf{F}_1$ und $\mathbf{F}_2$ inkremental beschränkt. Dann gilt $\xi_0(\mathbf{F}) \leq \xi_0(\mathbf{F}_1) \cdot \xi_0(\mathbf{F}_2)$.

Beweis:

Der Beweis ergibt sich direkt aus der Definition der inkrementalen Verstärkung.

2.3 Lineare Operatoren

Es seien $\mathbf{X}$ und $\mathbf{Y}$ zwei lineare Räume (beide reell oder beide komplex). Ein Operator $\mathbf{F}: \mathbf{X} \rightarrow \mathbf{Y}$ ist linear, wenn $\mathbf{F}(a_1 x_1 + a_2 x_2) = a_1 \mathbf{F} x_1 + a_2 \mathbf{F} x_2$ für alle $x_1, x_2 \in \mathbf{X}$ und $a_1, a_2 \in \mathbf{K}$ gilt. $\mathbf{K}$ ist dabei der Körper des linearen Raumes.

Man kann zeigen, daß Stetigkeit und Beschränktheit eines linearen Operators äquivalente Begriffe sind, wenn $\mathbf{X}$ und $\mathbf{Y}$ normierte lineare Räume sind [4].

2.3.1 Die Norm eines linearen, beschränkten Operators

Es seien $\mathbf{X}$ und $\mathbf{Y}$ zwei normierte lineare Räume. Der lineare Operator $\mathbf{F}$ bilde $\mathbf{X}$ in $\mathbf{Y}$ ab und sei beschränkt. Dann existiert die Verstärkung $g(\mathbf{F})$, und wir nennen sie die Norm $\|\mathbf{F}\|$ des linearen Operators $\mathbf{F}$.

Nach der Definition gilt

$$\|F\| = \sup_{\substack{x \in X \\ x \neq \theta}} \frac{\|Fx\|}{\|x\|}.$$

Eine gleichwertige Definition ist

$$\|F\| = \sup_{\substack{x \in X \\ \|x\|=1}} \|Fx\|,$$

da

$$\sup_{\substack{x \in X \\ x \neq \theta}} \frac{\|Fx\|}{\|x\|} = \sup_{\substack{x \in X \\ x \neq \theta}} \left\| F \frac{x}{\|x\|} \right\| = \sup_{\substack{x \in X \\ \|x\|=1}} \|Fx\|$$

gilt. Für alle linearen Operatoren F ist $F\theta = \theta$ und damit gilt $\|Fx\| \leq \|F\| \cdot \|x\|$ für alle $x \in X$. Ferner ist

$$\|F\| = g(F) = g_0(F).$$

Die Menge $L(X, Y)$ der linearen und beschränkten Operatoren $F: X \rightarrow Y$ bildet offenbar einen linearen Raum, wenn man als Nullelement den Nulloperator O wählt, der jedes Element $x \in X$ in das Nullelement von Y abbildet. Wie man leicht verifiziert, hat die Norm linearer beschränkter Operatoren die folgenden Eigenschaften, auf Grund derer $L(X, Y)$ zu einem normierten Raum wird:

- (i) $\|F\| = 0 \iff F = O,$
- (ii) $\|\lambda F\| = |\lambda| \cdot \|F\|$ für alle $\lambda \in K$, für alle $F \in L(X, Y)$
(K = Menge der reellen oder komplexen Zahlen),
- (iii) $\|F_1 + F_2\| \leq \|F_1\| + \|F_2\|$ für alle $F_1, F_2 \in L(X, Y).$

2.4 Nichtlineare Operatoren

Nichtlineare Operatoren sind Operatoren, die nicht notwendig linear sind. Lineare Operatoren sind also ein Spezialfall der nichtlinearen Operatoren. Die Wortschöpfung ist irreführend, sie ist aber üblich und wird deshalb beibehalten.

2.5 Die Inverse eines Operators

Der Operator $\mathbf{F}$ bilde die Menge $\mathbf{X}$ auf die Menge $\mathbf{Y}$ ab. (Die Abbildung ist also surjektiv.) Ferner sei die Abbildung injektiv (eindeutig), d.h. aus $\mathbf{F}x_1 = \mathbf{F}x_2$ mit $x_1, x_2 \in \mathbf{X}$ folgt immer $x_1 = x_2$.

Unter diesen Voraussetzungen ist die Abbildung bijektiv, d.h. zu jedem $y \in \mathbf{Y}$ gibt es genau ein $x \in \mathbf{X}$, so daß $y = \mathbf{F}x$ gilt. Anders ausgedrückt: Es existiert ein Operator $\mathbf{F}^{-1}$, die Inverse von $\mathbf{F}$ genannt, der $\mathbf{Y}$ auf $\mathbf{X}$ abbildet, derart, daß $\mathbf{F}^{-1}\mathbf{F} = \mathbf{I}_x$ und $\mathbf{F}\mathbf{F}^{-1} = \mathbf{I}_y$ gilt, wobei $\mathbf{I}_x$ und $\mathbf{I}_y$ die identischen Operatoren bezüglich $\mathbf{X}$ und $\mathbf{Y}$ sind. Folgende Sätze können leicht bewiesen werden :

Satz 7: Gegeben seien die Operatoren $\mathbf{G}:\mathbf{X} \rightarrow \mathbf{Y}$ und $\mathbf{F}:\mathbf{Y} \rightarrow \mathbf{Z}$. Die Inversen von $\mathbf{G}$ und $\mathbf{F}$ seien existent. Dann gilt $(\mathbf{F}\mathbf{G})^{-1} = \mathbf{G}^{-1}\mathbf{F}^{-1}$.

Satz 8: Falls ein linearer Operator eine Inverse besitzt, ist diese ebenfalls linear.

Satz 9: $\mathbf{X}$ und $\mathbf{Y}$ seien lineare normierte Räume. $\mathbf{F}$ sei ein linearer Operator, der $\mathbf{X}$ auf $\mathbf{Y}$ abbildet. Ferner gälte für alle $x \in \mathbf{X}$ und ein bestimmtes $m > 0$ $\|\mathbf{F}x\| \geq m \cdot \|x\|$. Dann existiert $\mathbf{F}^{-1}$ als beschränkter linearer Operator, und es gilt $\|\mathbf{F}^{-1}\| \leq \frac{1}{m}$.

Beweis:

Die Abbildung $\mathbf{F}$ ist eineindeutig, da aus $x_1 \neq x_2$

$$\|\mathbf{F}x_1 - \mathbf{F}x_2\| = \|\mathbf{F}(x_1 - x_2)\| \geq m\|x_1 - x_2\| > 0$$

und damit $\mathbf{F}x_1 \neq \mathbf{F}x_2$ folgt. Da ferner nach Voraussetzung $\mathbf{F}$ eine Abbildung von $\mathbf{X}$ auf $\mathbf{Y}$ ist, existiert $\mathbf{F}^{-1}$. Weiter gilt mit $y = \mathbf{F}x$

$$\|y\| = \|\mathbf{F}x\| \geq m\|x\| = m\|\mathbf{F}^{-1}y\|$$

und damit

$$\|\mathbf{F}^{-1}y\| \leq \frac{1}{m}\|y\|$$

für alle $y \in \mathbf{Y}$.

q.e.d.

Von besonderer Bedeutung ist der folgende tiefliegende Satz, dessen relativ schwieriger Beweis z.B. in [7] nachgelesen werden kann.

Satz 10: $\mathbf{X}$ und $\mathbf{Y}$ seien Banachräume. Der lineare und beschränkte Operator $\mathbf{F}: \mathbf{X} \longrightarrow \mathbf{Y}$ besitze die Inverse $\mathbf{F}^{-1}$. Dann ist $\mathbf{F}^{-1}$ linear und beschränkt.

Mit Hilfe des Banachschen Fixpunktsatzes (siehe nächster Abschnitt) kann nachstehender Satz bewiesen werden [4]:

Satz 11: $\mathbf{X}$ sei ein Banachraum. Der lineare und beschränkte Operator $\mathbf{F}: \mathbf{X} \longrightarrow \mathbf{X}$ sei eine Kontraktion, d.h. es ist $\gamma = \|\mathbf{F}\| < 1$. Dann hat der Operator $(\mathbf{I} - \mathbf{F})$ eine beschränkte lineare Inverse, und es gilt

$$\|(\mathbf{I} - \mathbf{F})^{-1}\| \leq \frac{1}{1 - \gamma}.$$

2.6 Operatoren und Systeme

Technisch-physikalische, aber auch chemische, biologische und ökonomische Zusammenhänge und Gesetzmäßigkeiten können häufig durch Systemblockschaltbilder beschrieben werden. Die einzelnen Systeme einer solchen Anordnung haben dabei modellartigen Charakter und kennzeichnen Eigenschaften und Einflüsse von untergeordneten Zusammenhängen auf das Gesamtsystem. Diese Zusammenhänge können dabei häufig als Abbildungen von Zeitfunktionen in ebensolche formuliert werden. Die angemessene mathematische Beschreibung solcher Systeme erfolgt also durch Operatoren, die einen Funktionenraum in einen im allgemeinen anderen Funktionenraum abbilden, und die Elemente dieser Funktionenräume sind Funktionen der Zeit t (siehe Abb. (1)).

Im folgenden werden einige in der Praxis häufig auftretende System- bzw. Operatoreigenschaften definiert, wobei der Beobachtungszeitraum gleich der Menge der reellen Zahlen $\mathbf{R}$ sein soll (siehe [22]).

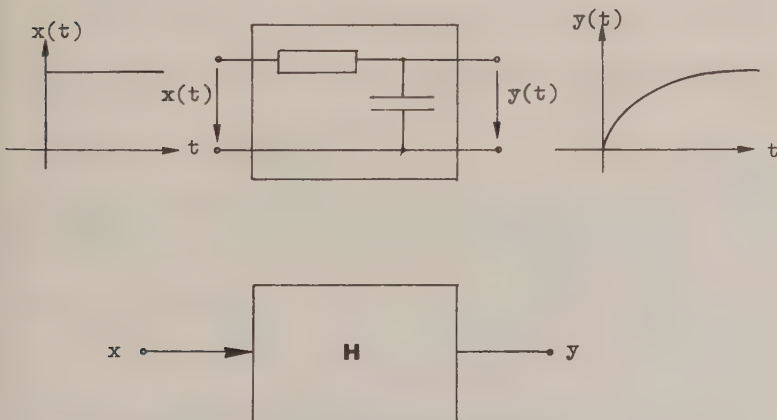


Abb. (1)

2.6.1 Zeitinvariante Systeme

Die Räume $\mathbf{X}$ und $\mathbf{Y}$ seien Funktionenräume, deren Elemente Funktionen der Zeit t , definiert auf $\mathbf{R}$, sind.

Das durch den Operator $\mathbf{F}:\mathbf{X}\rightarrow\mathbf{Y}$ beschriebene System heißt zeitinvariant, falls für alle $x\in\mathbf{X}$ mit $y=\mathbf{F}x$ auch $\hat{y}=\mathbf{F}\hat{x}$ gilt, wobei $\hat{x}$ und $\hat{y}$ aus x und y durch eine beliebige Verschiebung der Zeitachse entstehen. Das bedeutet:

Für alle $x\in\mathbf{X}$ mit

$$y(t) = \mathbf{F}x(t) \quad \text{für alle } t\in\mathbf{R}$$

gilt für ein fest gewähltes (aber beliebiges) $\tau\in\mathbf{R}$

$$y(t+\tau) = \mathbf{F}x(t+\tau) \quad \text{für alle } t\in\mathbf{R}$$

(siehe Abb. (2)).

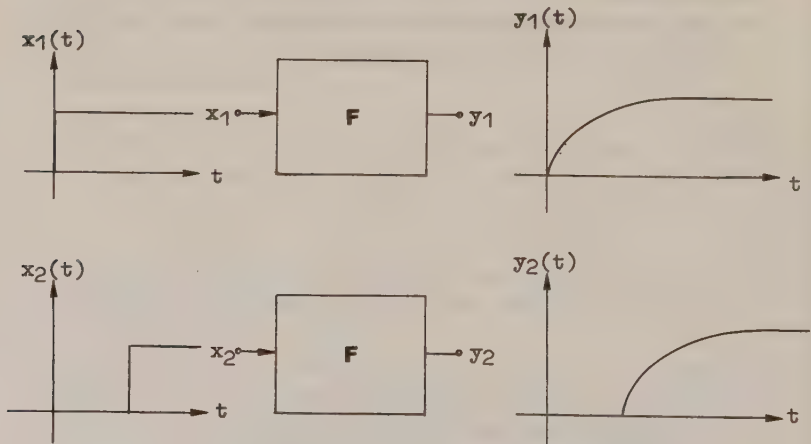


Abb. (2)

Dabei müssen die Funktionenräume $\mathbf{X}$ und $\mathbf{Y}$ dergestalt sein, daß auch die verschobenen Funktionen in ihnen enthalten sind.

2.6.2 Gedächtnisfreie Systeme

Die Räume $\mathbf{X}$ und $\mathbf{Y}$ seien Funktionenräume, deren Elemente Funktionen der Zeit t , definiert auf $\mathbf{R}$, sind.

Das durch den Operator $\mathbf{F}: \mathbf{X} \longrightarrow \mathbf{Y}$ beschriebene System heißt gedächtnisfrei, wenn für alle $x \in \mathbf{X}$ und alle $t_0 \in \mathbf{R}$ der Funktionswert von $\mathbf{F}x$ zum Zeitpunkt $t = t_0$ nur vom Funktionswert von x zum Zeitpunkt $t = t_0$ abhängt.

2.6.3 Kausale Systeme

Die Räume $\mathbf{X}$ und $\mathbf{Y}$ seien Funktionenräume, deren Elemente Funktionen der Zeit t , definiert auf $\mathbf{R}$, sind.

Die Definition der Kausalität eines Operators kann mittels eines Projektionsoperators $\mathbf{P}_T$, der durch

$$\mathbf{P}_T x(t) = \begin{cases} x(t) & \text{für } t \leq T \\ 0 & \text{für } t > T \end{cases}$$

gegeben ist, erfolgen:

Der Operator $\mathbf{F}: \mathbf{X} \longrightarrow \mathbf{Y}$ ist genau dann kausal, wenn für alle $T \in \mathbf{R}$

$$\mathbf{P}_T \mathbf{F} = \mathbf{P}_T \mathbf{F} \mathbf{P}_T$$

gilt. (Dabei müssen die Räume $\mathbf{X}$ bzw. $\mathbf{Y}$ dergestalt sein, daß für alle $x \in \mathbf{X}$ bzw. $y \in \mathbf{Y}$ auch $\mathbf{P}_T x \in \mathbf{X}$ bzw. $\mathbf{P}_T y \in \mathbf{Y}$ ist.)

3. Fixpunkte von Operatoren

3.1 Fixpunkte

Ein Operator F , der eine Menge X in sich abbildet, hat einen Fixpunkt $\hat{x} \in X$, wenn $\hat{x} = F \hat{x}$ gilt. Der Banachsche Fixpunktsatz gibt uns die Möglichkeit, unter gewissen Bedingungen die Existenz eines Fixpunktes zu beweisen. Dieser Satz hat für uns große Bedeutung, da in der Theorie der Rückkopplungssysteme häufig Operatorgleichungen auftreten, deren Lösungen als Fixpunkte von Operatoren aufgefaßt werden können.

3.2 Kontraktionen

Es sei X ein vollständiger metrischer Raum, der eine abgeschlossene Menge Ω enthalte. (Eine Teilmenge $\Omega \subset X$ eines metrischen Raumes heißt abgeschlossen, wenn der Grenzwert jeder konvergenten Folge $\{x_n\}$ mit $x_n \in \Omega$ in Ω liegt.)

F sei ein inkremental beschränkter Operator, der Ω in sich abbildet. Dann gilt nach Abschnitt I.2.2 für alle $x, y \in \Omega$

$$d(Fx, Fy) \leq \gamma \cdot d(x, y),$$

wobei γ die inkrementale Verstärkung von F bezüglich Ω ist. Falls $\gamma < 1$ ist, nennt man den Operator F eine Kontraktion, und γ heißt Kontraktionskonstante.

3.3 Der Banachsche Fixpunktsatz

Wir kommen jetzt zu einem Satz, der die Grundlage für die im Kapitel II zu behandelnde Systemtheorie darstellt:

Satz 12: Es sei X ein vollständiger metrischer Raum, der eine abgeschlossene Menge Ω enthalte. Der Operator $F: \Omega \rightarrow \Omega$ sei eine Kontraktion mit der Kontraktionskonstanten γ . Dann existiert in Ω ein eindeutiger Fixpunkt $\hat{x}$ von F , und es gilt

$$\hat{x} = \lim_{\mu \rightarrow \infty} x_{\mu} = \lim_{\mu \rightarrow \infty} F^{\mu} x_0$$

wobei $x_{\mu+1} = Fx_{\mu}$ für $\mu = 0, 1, 2, \dots$ ist.
(F^{μ} bedeutet die μ -fache Hintereinanderschaltung des Operators F .) x_0 ist dabei irgendein Element aus Ω . Ferner gilt

$$d(\hat{x}, x_{\mu}) \leq \frac{\gamma^{\mu}}{1-\gamma} d(x_1, x_0). \quad (3)$$

Beweis:

Für zwei beliebige nichtnegative ganze Zahlen μ, ν gilt mit $x_0 \in \Omega$

$$\begin{aligned} d(y_{\mu+\nu}, x_{\mu}) &= d(F^{\mu+\nu} x_0, F^{\mu} x_0) \\ &\leq \gamma^{\mu} d(F^{\nu} x_0, x_0) \\ &= \gamma^{\mu} d(x_{\nu}, x_0) \\ &\leq \gamma^{\mu} \{d(x_{\nu}, x_{\nu-1}) + d(x_{\nu-1}, x_{\nu-2}) + \dots + d(x_1, x_0)\} \\ &\leq \gamma^{\mu} \{ \gamma^{\nu-1} + \gamma^{\nu-2} + \dots + \gamma + 1 \} \cdot d(x_1, x_0) \\ &\leq \frac{\gamma^{\mu}}{1-\gamma} d(x_1, x_0). \end{aligned} \quad (4)$$

Aus $\lim_{\mu \rightarrow \infty} \gamma^\mu = 0$ folgt, daß $\{x_\mu\}$ eine Cauchy-Folge ist, und da X ein vollständiger Raum ist, konvergiert x_μ gegen ein $\hat{x} \in X$. Ferner ist jedes Folgenglied x_μ ein Element der abgeschlossenen Menge Ω und damit gilt $\hat{x} \in \Omega$. Da aus der Kontraktionseigenschaft eines Operators unmittelbar auch seine Stetigkeit folgt, gilt

$$\hat{x} = \lim_{\mu \rightarrow \infty} x_\mu = \lim_{\mu \rightarrow \infty} Fx_{\mu-1} = F(\lim_{\mu \rightarrow \infty} x_{\mu-1}) = F\hat{x}.$$

$\hat{x}$ ist also ein Fixpunkt des Operators F . Um zu zeigen, daß der Fixpunkt $\hat{x}$ eindeutig ist, nehmen wir an, daß es einen zweiten Fixpunkt x gibt. Dann ist

$$d(\hat{x}, x) = d(F\hat{x}, Fx) \leq \gamma \cdot d(\hat{x}, x),$$

und mit $\gamma < 1$ folgt $d(\hat{x}, x) = 0$ und schließlich $\hat{x} = x$. Die Ungl. (3) entsteht aus der Ungl. (4) für $v \rightarrow \infty$.

q.e.d.

In dieser Arbeit benötigen wir den Banachschen Fixpunktsatz nur in der folgenden speziellen Form:

Satz 13: X sei ein Banach-Raum. Der Operator $F: X \rightarrow X$ sei eine Kontraktion, d.h. es gibt ein $\gamma < 1$ derart, daß $\|Fx - Fy\| \leq \gamma \cdot \|x - y\|$ für alle $x, y \in X$ gilt. Dann existiert in X ein eindeutiger Fixpunkt $\hat{x}$, und es gilt

$$\hat{x} = \lim_{\mu \rightarrow \infty} x_\mu = \lim_{\mu \rightarrow \infty} F^\mu x_0,$$

wobei $x_{\mu+1} = Fx_\mu$ für $\mu = 0, 1, 2, \dots$ ist. x_0 ist dabei irgendein Element aus X . Ferner gilt

$$\|\hat{x} - x_\mu\| \leq \frac{\gamma^\mu}{1-\gamma} \|x_1 - x_0\|.$$

Der Banachsche Fixpunktsatz ist ein konstruktiver Existenzsatz. Es wird also nicht nur die Existenz eines Fixpunktes bewiesen, sondern auch eine Methode zu seiner Konstruktion angegeben, die durchaus zu einer praktischen Berechnung herangezogen werden kann. In der numerischen Mathematik nennt man Verfahren dieser Art Iterationsverfahren. Man geht von einem beliebigen Startpunkt $x_0 \in \Omega$ als einer nullten Näherung für den Fixpunkt aus und berechnet sukzessive verbesserte Näherungen nach der Vorschrift $x_{\mu+1} = Fx_{\mu}$ für $\mu = 0, 1, 2, \dots$, wobei der Fixpunktsatz die Konvergenz des Verfahrens garantiert.

In der Praxis wird man sich mit einer näherungsweise Berechnung des exakten Fixpunktes $\hat{x}$ zufrieden geben. Man wird versuchen, aus der Distanz zweier aufeinanderfolgender Näherungen $x_{\mu-1}$ und x_{μ} eine Schranke für den Fehler der Näherung x_{μ} zu ermitteln. Dazu betrachtet man Ungl. (3) für $\mu = 0$

$$d(\hat{x}, x_0) \leq \frac{1}{1-\gamma} d(x_1, x_0).$$

Wenn man nun die Näherung x_{μ} als nullte Näherung ansieht und $x_{\mu+1}$ als erste, folgt sofort

$$d(\hat{x}, x_{\mu}) \leq \frac{1}{1-\gamma} d(x_{\mu}, x_{\mu+1})$$

und daraus

$$d(\hat{x}, x_{\mu}) \leq \frac{\gamma}{1-\gamma} d(x_{\mu-1}, x_{\mu}).$$

Aus dieser Gleichung kann ein Abbruchkriterium für das Iterationsverfahren abgeleitet werden: Fordert man für den Fehler der μ -ten Näherung

$$d(\hat{x}, x_{\mu}) \leq \varepsilon, \tag{5}$$

so kann dies dadurch erreicht werden, daß die letzten beiden Näherungen die Bedingung

$$d(x_\mu, x_{\mu-1}) \leq \frac{1-\gamma}{\gamma} \epsilon \quad (6)$$

erfüllen.

Für $\gamma \rightarrow 1$ wird die Konvergenzgeschwindigkeit des Iterationsverfahrens auf Grund Ungl. (3) gering sein. In diesem Fall tritt eine weitere Schwierigkeit auf: Fordert man für den Fehler der μ -ten Iteration Ungl. (5), dann folgt aus Ungl. (6) mit $\gamma = 1-\delta$ die Bedingung

$$d(x_\mu, x_{\mu-1}) \leq \frac{\delta \epsilon}{1-\delta}$$

für die letzten beiden Iterationsschritte, welche auf Grund numerischer Gegebenheiten oft nicht zu erfüllen ist (Rundungsfehlerrauschen).

4. Fourier-Transformation und Faltungsoperatoren

Mit der Fourier-Transformation besitzen wir ein wichtiges funktionalanalytisches Hilfsmittel für die Beurteilung der Eigenschaften einer Klasse linearer Operatoren, insbesondere Faltungsoperatoren, die in der Theorie linearer Systeme von Bedeutung sind. Der Definitions- und Bildbereich dieser Operatoren seien dabei Funktionsräume, und die Elemente dieser Räume sollen Funktionen der reellen Zeit t sein. (Wir sprechen daher vom Zeitbereich.) Die Charakterisierung dieser Operatoren kann aber auch im Bereich der fouriertransformierten Funktionen (oder Frequenzbereich) erfolgen. Zeitfunktionen werden dabei durch Kleinbuchstaben und Frequenzfunktionen durch Großbuchstaben dargestellt.

Die den folgenden Betrachtungen zugrunde liegenden Funktionen entnehmen wir den Banachräumen L_1 und L_2 und weisen darauf hin, daß alle vorkommenden Integrale im Lebesgueschen Sinne zu verstehen sind. Falls über Matrizen oder Vektoren integriert wird, ist die Integration elementeweise zu verstehen.

Für eine beliebige $n \times m$ -Matrix M sei $\Lambda\{M\}$ die positive Quadratwurzel des größten Eigenwertes von $M^H M$, d.h.

$$\Lambda\{M\} = \sqrt{\lambda_{\max}(M^H M)}.$$

Zunächst sei auf eine wichtige Eigenschaft von L_2 -Funktionen hingewiesen:

Falls $f, g \in L_2(-\infty, \infty)$ gilt, ist das Produkt fg (definiert durch $fg(t) = f(t) \cdot g(t)$, für alle $t \in \mathbb{R}$) ein Element von $L_1(-\infty, \infty)$, und es erfüllt die Schwarzsche Ungleichung

$$\|fg\|_1 \leq \|f\|_2 \cdot \|g\|_2.$$

Die Indizes der Normen bezeichnen den L_p -Raum, auf den sie sich beziehen.

Bemerkung:

M^H bezeichnet die zur Matrix M konjugiert komplexe und transponierte Matrix. (Dies gilt natürlich auch für Vektoren.)

4.1 Fourier-Transformation von L_1 -Funktionen

Für $f \in L_1^n(-\infty, \infty)$ definieren wir die Fourier-Transformierte $F(j\omega)$ von $f(t)$ durch

$$F(j\omega) = \int_{-\infty}^{\infty} f(t) e^{-j\omega t} dt,$$

wobei $F(j\omega)$ für alle reellen ω existiert. Weitere Eigenschaften der Fourier-Transformierten sind [21] :

(i) $F_i(j\omega)$ ist gleichmäßig beschränkt, da

$$\sup_{\omega \in \mathbb{R}} |F_i(j\omega)| \leq \sup_{\omega \in \mathbb{R}} \int_{-\infty}^{\infty} |f_i(t) e^{-j\omega t}| dt \leq \int_{-\infty}^{\infty} |f_i(t)| dt = \|f_i\|$$

für $i = 1, \dots, n$ gilt.

(ii) $F_i(j\omega)$ und damit $F(j\omega)$ ist gleichmäßig stetig, da

$$|F_i(j\omega + h) - F_i(j\omega)| \leq \int_{-\infty}^{\infty} |e^{-jht} - 1| \cdot |f_i(t)| dt$$

gilt. Das Integral auf der rechten Seite konvergiert mit $h \rightarrow 0$ gegen Null. (Die Abschätzung ist unabhängig von ω .)

(iii) Nach dem Riemann-Lebesgue-Lemma gilt

$$\lim_{\omega \rightarrow \pm\infty} F(j\omega) = 0.$$

Oft stellt sich die Frage nach der Rücktransformation von $F(j\omega)$ in den Zeitbereich. Unter bestimmten einschränkenden Voraussetzungen gilt

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(j\omega) e^{j\omega t} d\omega.$$

Hinreichende Bedingungen für die Gültigkeit der (eindeutigen) Rücktransmutationsformel sind z.B. [4] :

- (i) $f(t)$ ist stetig und
- (ii) $F(j\omega)$ ist über ω absolut integrabel, d.h. mit $\hat{F}(\omega) = F(j\omega)$ gilt $\hat{F} \in L_1^n(-\infty, \infty)$.

Die Behandlung der Rücktransformation soll an dieser Stelle nicht weiter verfolgt werden, da sie in dieser Arbeit nur für L_2 -Funktionen von Interesse ist.

Bemerkung:

Für $f \in L_{1r}^n(-\infty, \infty)$ gilt $F(j\omega) = F^*(-j\omega)$ für alle $\omega \in \mathbb{R}$. Dabei bedeutet der Stern den konjugiert komplexen Wert.

4.2 Fourier-Transformation von L_2 -Funktionen

Die Fourier-Transformation einer Funktion $f \in L_2^n(-\infty, \infty)$ wird nach Plancherel durch

$$F(j\omega) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T}^T f(t) e^{-j\omega t} dt$$

definiert, wobei gezeigt werden kann, daß dieser Grenzwert existiert. Das Zeichen l.i.m. bedeutet Limes im Mittel und ist durch

$$\|\hat{F} - \hat{F}_T\|_2 \longrightarrow 0 \quad \text{für } T \longrightarrow \infty$$

mit

$$\hat{F}(\omega) = F(j\omega)$$

und

$$\hat{F}_T(\omega) = F_T(j\omega) = \int_{-T}^T f(t) e^{-j\omega t} dt$$

definiert.

Es seien f und g Elemente von $L_2^n(-\infty, \infty)$. Die entsprechenden Fourier-Transformierten seien $F(j\omega) = \hat{F}(\omega)$ und $G(j\omega) = \hat{G}(\omega)$. Dann gilt (siehe Bemerkung)[21,24]

$$(i) \quad \hat{F}, \hat{G} \in L_2^n(-\infty, \infty),$$

$$(ii) \quad \int_{-\infty}^{\infty} f^H(t)g(t)dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} F^H(j\omega)G(j\omega)d\omega$$

(Identität von Plancherel),

$$(iii) \quad \|f\|_2 = \left\{ \frac{1}{2\pi} \int_{-\infty}^{\infty} F^H(j\omega)F(j\omega)d\omega \right\}^{1/2}$$

(Das ist ein Spezialfall von (ii) und heißt Parseval'sches Theorem.)

$$(iv) \quad f(t) = \lim_{N \rightarrow \infty} \frac{1}{2\pi} \int_{-N}^N F(j\omega)e^{j\omega t} d\omega$$

(Das ist die eindeutige Rücktransformation in den Originalbereich.)

Eine weitere wichtige Eigenschaft ist:

(v) Falls mit $F(j\omega) = \hat{F}(\omega)$ und $\hat{F} \in L_2^n(-\infty, \infty)$

$$F(j\omega) = F^*(-j\omega)$$

für fast alle $\omega \in \mathbf{R}$ gilt, ist $F(j\omega)$ die Fourier-Transformierte einer eindeutigen Zeitfunktion $f \in L_{2\mathbf{R}}^n(-\infty, \infty)$, welche gemäß (iv) ermittelt werden kann.

Bemerkung 1:

Auf Grund der Schwarzschen Ungleichung und (i) existieren die Integrale von (ii).

Bemerkung 2:

Für $f \in L_{2\mathbf{R}}^n(-\infty, \infty) \cap L_{1\mathbf{R}}^n(-\infty, \infty)$ ist die L_2 -Fourier-Transformierte von f gleich der L_1 -Fourier-Transformierten und kann deshalb durch ein gewöhnliches Integral ermittelt werden:

$$F(j\omega) = \int_{-\infty}^{\infty} f(t) e^{-j\omega t} dt.$$

4.3 Lineare Operatoren $L_{2\mathbf{R}}^n(-\infty, \infty) \longrightarrow L_{2\mathbf{R}}^n(-\infty, \infty)$

Der folgende Satz bezieht sich auf eine Klasse linearer Operatoren, die im Frequenzbereich definiert sind [14] .

Satz 14: Gegeben sei eine meßbare, $n \times m$ -matrixwertige Funktion $H(j\omega)$ (mit im allgemeinen komplexwertigen Elementen) der reellen Variablen ω mit dem Definitionsbereich $\mathbf{R}$. Ferner gelte

- (i) $H(j\omega) = H^*(-j\omega)$ für fast alle $\omega \in \mathbf{R}$ und
(ii) $\text{ess sup}_{\omega \in \mathbf{R}} \{H(j\omega)\} < \infty$.

Der Operator $\mathbf{H}$ sei mit $f = \mathbf{H}g$ und $g \in \mathbf{L}_{2\mathbf{r}}^{\mathbf{n}}(-\infty, \infty)$ durch

$$F(j\omega) = H(j\omega)G(j\omega) \quad (7)$$

für fast alle $\omega \in \mathbf{R}$ definiert, wobei $G(j\omega)$ die $\mathbf{L}_2$ -Fourier-Transformierte von g und f die (wegen (ii) existente) $\mathbf{L}_2$ -Fourier-Transformierte von $F(j\omega)$ ist. Dann bildet der Operator $\mathbf{H}$ den Raum $\mathbf{L}_{2\mathbf{r}}^{\mathbf{n}}(-\infty, \infty)$ in den Raum $\mathbf{L}_{2\mathbf{r}}^{\mathbf{m}}(-\infty, \infty)$ ab. Er ist linear, beschränkt, und es gilt

$$\|\mathbf{H}\| = \text{ess sup}_{\omega \in \mathbf{R}} \{H(j\omega)\}.$$

Beweis:

Da $H(j\omega)$ und $G(j\omega)$ meßbare Funktionen sind, ist auch $F(j\omega)$ meßbar. Wegen $g \in \mathbf{L}_{2\mathbf{r}}^{\mathbf{n}}(-\infty, \infty)$ ist mit $\hat{G}(\omega) = G(j\omega)$ auch $\hat{G} \in \mathbf{L}_2^{\mathbf{n}}(-\infty, \infty)$, und es gilt

$$\begin{aligned} \int_{-\infty}^{\infty} F^{\mathbf{H}}(j\omega)F(j\omega)d\omega &= \int_{-\infty}^{\infty} G^{\mathbf{H}}(j\omega)H^{\mathbf{H}}(j\omega)H(j\omega)G(j\omega)d\omega \\ &\leq \int_{-\infty}^{\infty} \Lambda^2\{H(j\omega)\}G^{\mathbf{H}}(j\omega)G(j\omega)d\omega \quad (8) \\ &\leq \left\{ \text{ess sup}_{\omega \in \mathbf{R}} \{H(j\omega)\} \right\}^2 \int_{-\infty}^{\infty} G^{\mathbf{H}}(j\omega)G(\omega)d\omega < \infty. \end{aligned}$$

Daraus folgt mit $\hat{F}(\omega) = F(j\omega)$ zunächst $\hat{F} \in \mathbf{L}_2^{\mathbf{n}}(-\infty, \infty)$. Wegen (v) von Abschnitt I.4.2 existiert also für jedes $g \in \mathbf{L}_{2\mathbf{r}}^{\mathbf{n}}(-\infty, \infty)$ ein eindeutiges $f \in \mathbf{L}_2^{\mathbf{n}}(-\infty, \infty)$, für das Gl. (7) erfüllt ist. Auf Grund Bemerkung 2 von Abschnitt I.4.2 ist $G(j\omega) = G^*(-j\omega)$ für fast alle $\omega \in \mathbf{R}$ und damit

$$\begin{aligned} F(j\omega) &= H(j\omega)G(j\omega) = H^*(-j\omega)G^*(-j\omega) \\ &= (H(-j\omega)G(-j\omega))^* = F^*(-j\omega) \end{aligned}$$

für fast alle $\omega \in \mathbf{R}$, womit gezeigt ist, daß

$$\mathbf{H}: \mathbf{L}_{2\mathbf{r}}^n(-\infty, \infty) \longrightarrow \mathbf{L}_{2\mathbf{r}}^m(-\infty, \infty)$$

gilt. Der Nachweis der Linearität von $\mathbf{H}$ ist verhältnismäßig leicht zu führen.

Aus Gl. (8) folgt mit (iii) von Abschnitt I.4.2

$$\|f\| \leq \operatorname{ess\,sup}_{\omega \in \mathbf{R}} \Lambda \{H(j\omega)\} \cdot \|g\|$$

für alle $g \in \mathbf{L}_{2\mathbf{r}}^n(-\infty, \infty)$, und damit ist $\mathbf{H}$ beschränkt. Für die Norm des Operators $\mathbf{H}$ gilt also

$$\|\mathbf{H}\| \leq \operatorname{ess\,sup}_{\omega \in \mathbf{R}} \Lambda \{H(j\omega)\}.$$

Um zu beweisen, daß $\operatorname{ess\,sup}_{\omega} \Lambda \{H(j\omega)\}$ die kleinste obere Schranke der Abbildung ist, konstruieren wir für jedes $\epsilon > 0$ ein $g_\epsilon \in \mathbf{L}_{2\mathbf{r}}^n(-\infty, \infty)$ derart, daß mit $f_\epsilon = \mathbf{H}g_\epsilon$

$$\|f_\epsilon\| > (\operatorname{ess\,sup}_{\omega \in \mathbf{R}} \Lambda \{H(j\omega)\} - \epsilon)^2 \cdot \|g_\epsilon\|$$

gilt. Dazu wählen wir zunächst $G_\epsilon(j\omega)$ so, daß $\hat{G}_\epsilon \in \mathbf{L}_2^n(-\infty, \infty)$ ist und

$$\int_{-\infty}^{\infty} F_\epsilon^H(j\omega) F_\epsilon(j\omega) d\omega = \int_{-\infty}^{\infty} \Lambda^2 \{H(j\omega)\} G_\epsilon^H(j\omega) G_\epsilon(j\omega) d\omega \quad (9)$$

gilt. Dies läßt sich immer dadurch erreichen, daß man für fast alle $\omega \in \mathbf{R}$ $G_\epsilon(j\omega)$ als Eigenvektor zum größten Eigenwert von $H^H(j\omega)H(j\omega)$ wählt. Wie man leicht erkennt, kann die Funktion $G_\epsilon(j\omega)$ so festgelegt werden, daß sowohl Gl. (9) als auch

$$G_{\epsilon}(j\omega) = G_{\epsilon}^{*}(-j\omega) \quad (10)$$

für fast alle $\omega \in \mathbf{R}$ erfüllt wird, da mit $G_{\epsilon}(j\omega)$ als Eigenvektor zum größten Eigenwert von $H^H(j\omega)H(j\omega)$ automatisch $G_{\epsilon}^{*}(j\omega)$ einen Eigenvektor zum größten Eigenwert von $H^H(-j\omega)H(-j\omega)$ darstellt (wegen der Gleichheit der Eigenwerte beider Matrizen):

$$H^H(j\omega)H(j\omega)G_{\epsilon}(j\omega) = \lambda G_{\epsilon}(j\omega)$$

$$H^{H*}(j\omega)H^{*}(j\omega)G_{\epsilon}^{*}(j\omega) = \lambda G_{\epsilon}^{*}(j\omega)$$

$$H^H(-j\omega)H(-j\omega)G_{\epsilon}^{*}(j\omega) = \lambda G_{\epsilon}^{*}(j\omega).$$

Jetzt wählen wir $G_{\epsilon}(j\omega)$ derart, daß folgende Bedingungen erfüllt sind:

$$(i) \quad \hat{G}_{\epsilon} \in L_2^n(-\infty, \infty),$$

$$(ii) \quad \hat{G}_{\epsilon} \neq 0,$$

$$(iii) \quad \int_{-\infty}^{\infty} \Lambda^2\{H(j\omega)\} G_{\epsilon}^H(j\omega) G_{\epsilon}(j\omega) d\omega > (\text{ess sup}_{\omega \in \mathbf{R}} \Lambda\{H(j\omega)\} - \epsilon)^2 \int_{-\infty}^{\infty} G_{\epsilon}^H(j\omega) G_{\epsilon}(j\omega) d\omega$$

Das kann ohne Verletzung von Gl. (9) und Gl. (10) immer dadurch sichergestellt werden, daß man $G_{\epsilon}(j\omega)$ nur dort vom Nullvektor verschieden wählt, wo

$$\Lambda\{H(j\omega)\} > \text{ess sup}_{\omega \in \mathbf{R}} \Lambda\{H(j\omega)\} - \epsilon \quad (11)$$

gilt. Diese Konstruktion ist möglich, da die Gl. (11) nach der Definition des essentiellen Supremums auf einer Menge Δ mit dem Maß $m(\Delta) > 0$ erfüllt wird. Damit haben wir eine Vektorfunktion $G_{\epsilon}(j\omega)$ konstruiert, für die

$$\int_{-\infty}^{\infty} F_{\epsilon}^H(j\omega) F_{\epsilon}(j\omega) d\omega > (\text{ess sup}_{\omega \in \mathbf{R}} \Lambda\{H(j\omega)\} - \epsilon)^2 \int_{-\infty}^{\infty} G_{\epsilon}^H(j\omega) G_{\epsilon}(j\omega) d\omega$$

gilt und deren zugehörige Zeitfunktion ein Element des Raumes $L_{2r}^n(-\infty, \infty)$ ist. Nach Anwendung des Parsevalschen Theorems gilt

$$\|f_\epsilon\| > (\text{ess sup}_{\omega \in \mathbf{R}} \Lambda\{H(j\omega)\} - \epsilon) \|g_\epsilon\|.$$

Damit ist gezeigt, daß kein

$$\text{ess sup}_{\omega \in \mathbf{R}} \Lambda\{H(j\omega)\} - \epsilon$$

mit $\epsilon > 0$ obere Schranke der Abbildung ist, und folglich gilt

$$\|H\| = \text{ess sup}_{\omega \in \mathbf{R}} \Lambda\{H(j\omega)\}.$$

Bemerkung:

Im Falle $n = 1$ gilt offenbar

$$\|H\| = \text{ess sup}_{\omega \in \mathbf{R}} |H(j\omega)|.$$

4.4 Lineare Operatoren vom Faltungstyp

Betrachtet werden lineare Operatoren H , die mit $f = Hg$ durch folgende Faltungsbeziehung definiert sind:

$$f(t) = h_0 g(t) + \int_{-\infty}^{\infty} h(t-\tau) g(\tau) d\tau \quad \text{für alle } t \in \mathbf{R}.$$

Dabei sei h_0 eine reelle $n \times m$ -Matrix, und h sei eine $n \times m$ -matrixwertige Funktion. Alle Elemente von h seien im Raum $L_{1r}(-\infty, \infty)$ enthalten, und wir schreiben dafür

$$h \in M_{1r}^{n,m}(-\infty, \infty).$$

Abbildungen dieser Art treten in der Theorie linearer Systeme mit n Eingängen und m Ausgängen häufig auf, und man spricht dann von der Erregung g und Reaktion f des linearen Systems. Ferner nennt man die verallgemeinerte Matrixfunktion

$$\hat{h}(t) = h_0 \delta(t) + h(t)$$

die Matrix der Impulsantworten [22] .

Die Menge aller Faltungsoperatoren gemäß obigen Vereinbarungen nennen wir $F^{n,m}$. Wir wollen im folgenden diese Faltungsoperatoren auf Elemente des Raumes $L_{2r}^n(-\infty, \infty)$ anwenden.

Es kann gezeigt werden, daß mit $H \in F^{n,m}$

$$H : L_{2r}^n(-\infty, \infty) \longrightarrow L_{2r}^m(-\infty, \infty)$$

gilt, wobei die Abbildung $f = Hg$ im Frequenzbereich durch

$$F(j\omega) = H(j\omega)G(j\omega)$$

für fast alle $\omega \in \mathbf{R}$ definiert ist. Dabei sind $G(j\omega)$ und $F(j\omega)$ die L_2 -Fourier-Transformierten von g und f , und $H(j\omega)$ ist durch

$$H(j\omega) = h_0 + \int_{-\infty}^{\infty} h(t) e^{-j\omega t} dt \text{ für alle } \omega \in \mathbf{R}$$

gegeben. (Man nennt $H(j\omega)$ die Matrix der fouriertransformierten Impulsantworten.) Das Integral bezeichnet die L_1 -Fourier-Transformierte der Funktion h gemäß Abschnitt I.4.1. Jedes Element von $H(j\omega)$ ist also eine

gleichmäßig stetige Funktion der reellen Variablen ω ,
und es gilt

$$\lim_{\omega \rightarrow \pm \infty} H(j\omega) = h_0.$$

Ferner ist

$$H(j\omega) = H^*(-j\omega) \text{ für alle } \omega \in \mathbf{R}.$$

Da alle Elemente von $H(j\omega)$ gleichmäßig beschränkt
sind, gilt

$$\sup_{\omega \in \mathbf{R}} \wedge \{H(j\omega)\} < \infty,$$

und wegen der gleichmäßigen Stetigkeit ist

$$\operatorname{ess\,sup}_{\omega \in \mathbf{R}} \wedge \{H(j\omega)\} = \sup_{\omega \in \mathbf{R}} \wedge \{H(j\omega)\} < \infty.$$

Damit sind alle Voraussetzungen von Satz 14 erfüllt,
und folglich sind alle Operatoren $H \in \mathbf{F}^{n,m}$ lineare, be-
schränkte Operatoren, die den Raum $\mathbf{L}_{2r}^n(-\infty, \infty)$ in den
Raum $\mathbf{L}_{2r}^m(-\infty, \infty)$ abbilden, und es gilt

$$\|H\| = \sup_{\omega \in \mathbf{R}} \wedge \{H(j\omega)\}.$$

Bemerkung:

Durch die Operatoren $H \in \mathbf{F}^{n,m}$ können fast alle Systeme
beschrieben werden, die linear, asymptotisch stabil,
zeitinvariant und dabei nicht notwendig kausal sind [22] .

II. THEORIE EINER KLASSE NICHTLINEARER SYSTEME

1. Einleitung

In diesem Kapitel betrachten wir die Operatorgleichungen

$$e = f + \mathbf{H}\mathbf{N}f$$

und

$$e = \mathbf{N}f + \mathbf{H}f.$$

Dabei sei $\mathbf{H}$ ein linearer und $\mathbf{N}$ ein nichtlinearer Operator, während e die vorgegebene Systemerregung, und f die gesuchte Systemreaktion darstelle. Diesen Gleichungen können die Systemblockschaltbilder (Abb. (1) und Abb. (2)) zugeordnet werden, wobei zu beachten ist, daß die Pfeile bei den Teilsystemen nicht notwendig eine Ursache, sondern die Abbildungsrichtung der Operatoren definieren.

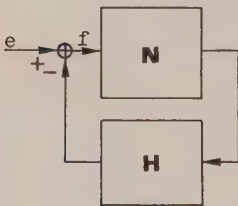


Abb. (1)

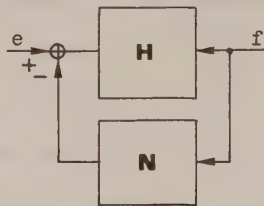


Abb. (2)

Systeme dieser Struktur treten bei der Untersuchung nichtlinearer physikalischer Gebilde wie z.B. in der Netzwerktheorie und in der Regelungstechnik immer dann auf, wenn es möglich ist, das gesamte Systemverhalten durch das Zusammenspiel eines linearen und eines nichtlinearen Systems zu beschreiben.

Unsere Aufgabe besteht darin, anhand des Banachschen Fixpunktsatzes Bedingungen für die eindeutige Lösbarkeit der Operatorgleichungen und Berechnungsverfahren zur Ermittlung der Lösungen anzugeben. Dabei werden wir zunächst von ganz allgemeinen Voraussetzungen ausgehen, um uns dann im Hinblick auf eine bestimmte technische Anwendung geeignet einzuschränken. Dies bedeutet konkret, daß wir eine allgemeine Theorie zur Analyse der Systeme gemäß Abb. (1) und Abb. (2) aufstellen, wobei wir uns z.B. bezüglich der Art der Systemerregung e bzw. der gesuchten Systemreaktion f nur insofern festlegen, als diese einem gewissen Banachraum angehören sollen. Danach folgt die Anwendung dieser Theorie auf die Analyse von Übergangsvorgängen einer Klasse technisch relevanter nichtlinearer Systeme der geforderten Struktur nach Abb. (1) bzw. Abb. (2). (Dabei wählen wir als Banachraum den Raum der quadratisch integrierbaren Funktionen, da mit Funktionen dieser Klasse sowohl wichtige praktische Fragestellungen als auch interessante theoretische Aussagen verknüpft sind.) Grundsätzlich ist die vorgestellte Theorie aber auch für die Analyse von periodischen oder zeitlich konstanten Vorgängen geeignet, wenn man den Banachraum entsprechend wählt. Immer wenn wir uns mit dem Banachraum auf einen Funktionenraum einschränken, soll es sich aus naheliegenden Gründen um Funktionen der Zeit t handeln.

In den Kapiteln III. und IV. wird die hier beschriebene Theorie praktisch angewandt.

2. Bedingungen für die Lösbarkeit der Operatorgleichungen

2.1 Der Banachraum $\mathbf{B}^n$

Der den folgenden Untersuchungen zugrundeliegende Banachraum $\mathbf{B}^n$ habe nachstehende Eigenschaften:

- (i) Jedes Element $h \in \mathbf{B}^n$ besteht aus n geordneten Komponenten h_i , d.h.

$$h = \begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_n \end{bmatrix},$$

wobei h genau dann ein Element von $\mathbf{B}^n$ ist, wenn jedes h_i ein Element des reellen Banachraumes $\mathbf{B}$ ist. (Man sagt auch $\mathbf{B}^n$ ist das n -fache kartesische Produkt von $\mathbf{B}$.)

- (ii) Addition und skalare Multiplikation sind im $\mathbf{B}^n$ komponentenweise erklärt.

- (iii) Der für alle $h \in \mathbf{B}^n$ durch

$$\mathbf{C}h = \begin{bmatrix} (\mathbf{C}h)_1 \\ (\mathbf{C}h)_2 \\ \vdots \\ (\mathbf{C}h)_n \end{bmatrix} = \begin{bmatrix} c_1 h_1 \\ c_2 h_2 \\ \vdots \\ c_n h_n \end{bmatrix}$$

mit vorgegebenem $(c_1, \dots, c_n)^T \in \mathbf{R}^n$ definierte lineare Operator $\mathbf{C}: \mathbf{B}^n \rightarrow \mathbf{B}^n$ hat die Norm

$$\|\mathbf{C}\| = \max_{i=1, \dots, n} |c_i|,$$

und es gilt

$$\|\mathbf{C}h\| \geq \min_{i=1, \dots, n} c_i \cdot \|h\| \quad \text{für alle } h \in \mathbf{B}^n.$$

Bemerkung 1:

Aus (i) und (ii) folgt, daß $\mathbf{B}^n$ ein reeller Banachraum ist.

Bemerkung 2:

Daß der in (iii) definierte Operator $\mathbf{C}$ den $\mathbf{B}^n$ in sich abbildet, ist eine Konsequenz aus (i). Wegen (i) und (ii) ist der Operator $\mathbf{C}$ linear.

Bemerkung 3:

Es kann leicht nachgewiesen werden, daß die Räume $\mathbf{L}_{pr}^n(-\infty, \infty)$, $\mathbf{R}^n$ und $\mathbf{P}_T^n$ die gesetzten Bedingungen erfüllen. (Dabei ist $\mathbf{P}_T^n$ der Banachraum der n -komponentigen periodischen Funktionen, die über eine Periode T quadratisch integrierbar sind.)

2.2 Die Operatorgleichungen

Wir betrachten Operatorgleichungen der Form

$$e = f + \mathbf{H}\mathbf{N}f \quad (1)$$

und

$$e = \mathbf{N}f + \mathbf{H}f, \quad (2)$$

wobei $\mathbf{H}$ und $\mathbf{N}$ Operatoren sein sollen, die den Raum $\mathbf{B}^n$ in sich abbilden. In den folgenden Abschnitten wird es unsere Aufgabe sein, Bedingungen für die eindeutige Lösbarkeit der Operatorgleichungen (1) und (2) bei vorgegebenem $e \in \mathbf{B}^n$ und Verfahren zur Ermittlung dieser Lösungen anzugeben.

Zunächst definieren wir die Transformationsoperatoren $\mathbf{C}$ und $\mathbf{T}$, deren Bedeutung erst später klar werden wird. Diese Operatoren werden für alle $h \in \mathbf{B}^n$ durch

$$(\mathbf{C}h)_i = c_i h_i \quad \text{für } i=1, \dots, n$$

und

$$(\mathbf{T}h)_i = t_i h_i \quad \text{für } i=1, \dots, n$$

mit fest gewählten Vektoren

$$\mathbf{c} = (c_1, c_2, \dots, c_n)^T \in \mathbb{R}^n$$

und

$$\mathbf{t} = (t_1, t_2, \dots, t_n)^T \in \mathbb{R}^n$$

erklärt.

Nach Abschnitt II.2.1 sind $\mathbf{C}$ und $\mathbf{T}$ beschränkte lineare Operatoren, die den $\mathbb{B}^n$ in sich abbilden.

Die in den Operatorgleichungen (1) und (2) vorkommenden Operatoren sollen folgende Eigenschaften haben:

- (i) $\mathbf{H}: \mathbb{B}^n \longrightarrow \mathbb{B}^n$ ist linear und beschränkt.
- (ii) $\mathbf{N}: \mathbb{B}^n \longrightarrow \mathbb{B}^n$ ist nichtlinear, und es existieren Vektoren $\alpha, \beta \in \mathbb{R}^n$ mit $\alpha_i < \beta_i$ für $i=1, \dots, n$ derart, daß für alle $h, k \in \mathbb{B}^n$ und alle Operatoren $\mathbf{C}$ und $\mathbf{T}$ folgende Gleichung gilt:

$$\|\mathbf{T}(\mathbf{N}-\mathbf{C})h - \mathbf{T}(\mathbf{N}-\mathbf{C})k\| \leq \eta(\mathbf{C}, \mathbf{T}) \cdot \|h - k\|. \quad (3)$$

Dabei ist

$$\eta(\mathbf{C}, \mathbf{T}) = \max_{i=1, \dots, n} (|t_i| \cdot \eta_i(c_i))$$

mit

$$\eta_i(c_i) = \max \{c_i - \alpha_i, \beta_i - c_i\}.$$

Mit der Bezeichnungsweise von Abschnitt I.2.2 soll also $g_0(\mathbf{T}(\mathbf{N}-\mathbf{C})) \leq \eta(\mathbf{C}, \mathbf{T})$ gelten.

Für den Fall, daß der Banachraum $\mathbf{B}^n$ gleich einem der $\mathbf{L}_{pr}^n$ -Räume ist, wird die Forderung (ii) an den nichtlinearen Operator $\mathbf{N}$ erfüllt, wenn dieser durch n nichtlineare Funktionen ψ_i definiert ist, welche auf jeweils eine Komponente h_i von $h \in \mathbf{L}_{pr}^n$ angewandt werden und dadurch die Abbildung definieren. Die Funktionen ψ_i müssen aber noch weitere Bedingungen erfüllen. (Die Zahlen α_i bzw. β_i stellen die Minimal- bzw. Maximalsteigungen der Funktionen ψ_i dar, falls diese differenzierbar sind.) Die genaue Behandlung dieses Spezialfalls erfolgt in Abschnitt II.2.3.

Bemerkung:

Für alle Vektoren $\hat{\alpha}, \hat{\beta} \in \mathbf{R}^n$ mit

$$\hat{\alpha}_i \leq \alpha_i, \beta_i \leq \hat{\beta}_i \quad \text{für } i=1, \dots, n \quad (4)$$

gilt

$$\eta(\mathbf{C}, \mathbf{T}) \leq \hat{\eta}(\mathbf{C}, \mathbf{T}),$$

wobei

$$\hat{\eta}(\mathbf{C}, \mathbf{T}) = \max_{i=1, \dots, n} (|t_i| \cdot \hat{\eta}_i(c_i))$$

und

$$\hat{\eta}_i(c_i) = \max \{ c_i - \hat{\alpha}_i, \hat{\beta}_i - c_i \}$$

ist.

Falls also die Gl. (3) für $\alpha, \beta \in \mathbf{R}^n$ erfüllt ist, gilt sie automatisch auch für alle $\hat{\alpha}, \hat{\beta} \in \mathbf{R}^n$, die der Gl. (4) genügen. Das diesen Vektoren $\hat{\alpha}, \hat{\beta}$ zugeordnete $\hat{\eta}(\mathbf{C}, \mathbf{T})$ stellt aber im allgemeinen eine schlechtere Abschätzung für $g_0(\mathbf{T}(\mathbf{N}-\mathbf{C}))$ dar.

2.3 Eine Klasse $\Psi(\alpha, \beta)$ nichtlinearer Operatoren, für die Gl. (3) gilt

Der Banachraum $\mathbf{B}^n$ sei einer der $\mathbf{L}_{pr}^n(-\infty, \infty)$ -Räume. Ferner seien $\psi_1(u, t), \psi_2(u, t), \dots, \psi_n(u, t)$ reellwertige Funktionen der reellen Variablen u und t , definiert für alle $u, t \in \mathbf{R}$, und es gelte:

- (i) Es existieren Vektoren $\alpha, \beta \in \mathbf{R}^n$ derart, daß mit $\alpha_i < \beta_i$ für $i=1, \dots, n$ die Ungleichung

$$\alpha_i \leq \frac{\psi_i(u, t) - \psi_i(v, t)}{u - v} \leq \beta_i$$

für $i=1, \dots, n$, alle $t \in \mathbf{R}$ und alle $u, v \in \mathbf{R}$ mit $u \neq v$ gelten.

- (ii) $\psi_i(0, t) = 0$ für $i=1, \dots, n$ und alle $t \in \mathbf{R}$.

- (iii) $\psi_i(u(\cdot), \cdot)$ ist eine meßbare Funktion der Zeit, falls $u(\cdot)$ eine meßbare Funktion ist (für $i=1, \dots, n$).

Am Rande sei erwähnt, daß die Bedingung (iii) für den Ingenieur ohne Bedeutung ist, da Funktionen, die diese Bedingungen nicht erfüllen, für die Anwendung ohne Interesse sind.

Der Operator $\mathbf{N}$ wird durch

$$(\mathbf{N}h)_i(t) = \psi_i(h_i(t), t) \quad \text{für } i=1, \dots, n$$

und alle $t \in \mathbf{R}, h \in \mathbf{B}^n$ definiert, wobei $(\mathbf{N}h)_i$ die i -te Komponente von $\mathbf{N}h$ ist, und wir schreiben

$$\begin{aligned} \mathbf{N}h(t) &= \psi(h(t), t) \\ &= (\psi_1(h_1(t), t), \psi_2(h_2(t), t), \dots, \psi_n(h_n(t), t))^T. \end{aligned}$$

Def. 1: Die Klasse der Operatoren, die obigen Bedingungen genügen, nennen wir $\Psi(\alpha, \beta)$.

Mit dieser Definition können wir folgenden Satz beweisen:

Satz 1: Es sei $N \in \Psi(\alpha, \beta)$. Dann ist $N: B^n \rightarrow B^n$, und es gilt

$$\|T(N-C)h - T(N-C)k\| \leq \eta(C, T) \cdot \|h - k\|$$

für alle $h, k \in B^n$ und beliebige Operatoren C und T der in Abschnitt II.2.2 zugelassenen Art.

Beweis:

Zunächst zeigen wir, daß $N: B^n \rightarrow B^n$ gilt. Nach Voraussetzung (i) gilt mit $v = 0$ und (ii)

$$\begin{aligned} |\psi_i(h_i(t), t)| &\leq \max \{ |\alpha_i|, |\beta_i| \} |h_i(t)| \\ &\leq \max_{i=1, \dots, n} \max \{ |\alpha_i|, |\beta_i| \} \cdot |h_i(t)|, \end{aligned}$$

und mit (iii) ist $\psi_i(h_i(\cdot), \cdot)$ meßbar, falls $h_i(\cdot)$ meßbar ist.

Es sei $h \in L_{pr}^n(-\infty, \infty)$. Für $1 \leq p < \infty$ gilt

$$\begin{aligned} \|Nh\| &= \left\{ \int_{-\infty}^{\infty} \sum_{i=1}^n |\psi_i(h_i(t), t)|^p dt \right\}^{1/p} \\ &\leq \max_{i=1, \dots, n} \max \{ |\alpha_i|, |\beta_i| \} \cdot \|h\| \end{aligned}$$

und damit $Nh \in L_{pr}^n(-\infty, \infty)$. Für $p = \infty$ ist

$$\begin{aligned} \|Nh\| &= \sum_{i=1}^n \operatorname{ess\,sup}_{t \in \mathbf{R}} |\psi_i(h_i(t), t)| \\ &\leq \max_{i=1, \dots, n} \max \{ |\alpha_i|, |\beta_i| \} \cdot \|h\| \end{aligned}$$

und damit $Nh \in L_{\infty}^n(-\infty, \infty)$.

Jetzt müssen wir noch die Gültigkeit der Gl. (3) nachweisen:
Aus der Voraussetzung (i) folgt

$$\begin{aligned} & (a_i - c_i)(h_i(t) - k_i(t)) \\ & \leq (\psi_i(h_i(t), t) - c_i h_i(t)) - (\psi_i(k_i(t), t) - c_i k_i(t)) \\ & \leq (\beta_i - c_i)(h_i(t) - k_i(t)) \end{aligned}$$

und daraus

$$\begin{aligned} & |(\psi_i(h_i(t), t) - c_i h_i(t)) - (\psi_i(k_i(t), t) - c_i k_i(t))| \\ & \leq \eta_i(c_i) \cdot |h_i(t) - k_i(t)| \end{aligned}$$

für $i=1, \dots, n$ und $h_i(t), k_i(t) \in \mathbf{R}$. Nach Multiplikation mit t_i gilt

$$\begin{aligned} & |t_i(\psi_i(h_i(t), t) - c_i h_i(t)) - t_i(\psi_i(k_i(t), t) - c_i k_i(t))| \\ & \leq |t_i| \cdot \eta_i(c_i) \cdot |h_i(t) - k_i(t)|, \end{aligned}$$

und wir kommen zu einer Ungleichung, die wir anschließend benötigen werden:

$$\begin{aligned} & |t_i(\psi_i(h_i(t), t) - c_i h_i(t)) - t_i(\psi_i(k_i(t), t) - c_i k_i(t))| \\ & \leq \eta(\mathbf{C}, \mathbf{T}) \cdot |h_i(t) - k_i(t)| \end{aligned}$$

für $i=1, \dots, n$ und alle $h_i(t), k_i(t), c_i, t_i \in \mathbf{R}$.

Es seien $h, k \in \mathbf{L}_{pr}^n(-\infty, \infty)$. Dann gilt für $1 \leq p < \infty$

$$\| \mathbf{T}(\mathbf{N} - \mathbf{C})h - \mathbf{T}(\mathbf{N} - \mathbf{C})k \| =$$

$$\left\{ \int_{-\infty}^{\infty} \sum_{i=1}^n |t_i(\psi_i(h_i(t), t) - c_i h_i(t)) - t_i(\psi_i(k_i(t), t) - c_i k_i(t))|^p dt \right\}^{1/p}$$

$$\leq \eta(\mathbf{C}, \mathbf{T}) \cdot \| h - k \|$$

und für $p = \infty$

$$\| \mathbf{T}(\mathbf{N}-\mathbf{C})\mathbf{h} - \mathbf{T}(\mathbf{N}-\mathbf{C})\mathbf{k} \| =$$

$$\sum_{i=1}^n \operatorname{ess\,sup}_{t \in \mathbf{R}} |t_i(\psi_i(h_i(t), t) - c_i h_i(t)) - t_i(\psi_i(k_i(t), t) - c_i k_i(t))|$$

$$\leq \eta(\mathbf{C}, \mathbf{T}) \|\mathbf{h} - \mathbf{k}\|.$$

q.e.d.

Bemerkung:

Aus der Definition der Klasse $\Psi(\alpha, \beta)$ ist leicht ersichtlich, daß es sich bei den Operatoren $\mathbf{N} \in \Psi(\alpha, \beta)$ um Abbildungen einer Vektorfunktion $\mathbf{h} \in \mathbf{L}_{\text{pr}}^n(-\infty, \infty)$ in eine Vektorfunktion $\mathbf{N}\mathbf{h} \in \mathbf{L}_{\text{pr}}^n(-\infty, \infty)$ handelt, wobei jede Komponente $h_i \in \mathbf{L}_{\text{pr}}(-\infty, \infty)$ von $\mathbf{h}$ durch einen nichtlinearen, zeitvarianten und gedächtnisfreien Suboperator in die entsprechende Komponente $(\mathbf{N}\mathbf{h})_i \in \mathbf{L}_{\text{pr}}(-\infty, \infty)$ von $\mathbf{h}$ abgebildet wird. Jeder dieser n Suboperatoren wird durch eine gewöhnliche Funktion $\psi_i(h_i(t), t)$ definiert, die in jedem Zeitpunkt nach oben und unten "steigungs"-beschränkt ist (bezüglich der unabhängigen Variablen $h_i(t)$). Dies soll anhand eines Beispiels gemäß Abb. (3) veranschaulicht werden.

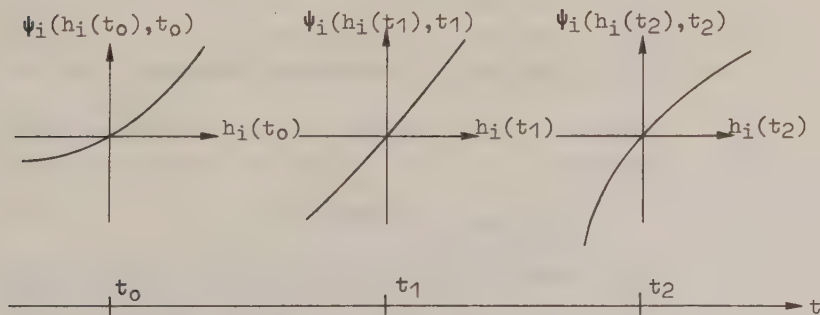


Abb. (3)

2.4 Allgemeine Bedingungen für die eindeutige Lösbarkeit der Operatorgleichungen

In diesem Kapitel wollen wir hinreichende Bedingungen für die eindeutige Lösbarkeit der Operatorgleichungen von Abschnitt II.2.2 angeben. Dabei wird der Banachsche Fixpunktsatz das wichtigste Beweishilfsmittel sein. Daraus ergibt sich dann auch ein Konstruktionsverfahren für die Ermittlung der gesuchten Lösung. Vor Anwendung dieses Satzes werden die Operatorgleichungen in neue Gleichungen transformiert, die die selben Lösungen wie die ursprünglichen Gleichungen aufweisen. Diese Transformation erfolgt, wie noch in Abschnitt II.2.4.3 gezeigt werden wird, im Hinblick auf eine Minimierung der Kontraktionsschranke.

2.4.1 Lösung der Gleichung $e = f + \mathbf{H}\mathbf{N}f$

Unter den Voraussetzungen der Abschnitte II.2.1 und II.2.2 kann der folgende Satz bewiesen werden:

Satz 2: Es seien die Operatoren $\mathbf{C}$ und $\mathbf{T}$ derart gegeben, daß folgende Bedingungen erfüllt sind:

- (i) $\mathbf{T}^{-1}$ existiert,
- (ii) $(\mathbf{I} + \mathbf{H}\mathbf{C})^{-1}$ existiert,
- (iii) $r(\mathbf{C}, \mathbf{T}) = \|(\mathbf{I} + \mathbf{H}\mathbf{C})^{-1}\mathbf{H}\mathbf{T}^{-1}\| \cdot \eta(\mathbf{C}, \mathbf{T}) < 1$.

Dann gibt es für jedes $e \in \mathbf{B}^n$ ein eindeutiges $f \in \mathbf{B}^n$, so daß

$$e = f + \mathbf{H}\mathbf{N}f \quad (4)$$

gilt. Ferner ist $f = \lim_{\mu \rightarrow \infty} f_\mu$, wobei

$$f_{\mu+1} = -(\mathbf{I} + \mathbf{H}\mathbf{C})^{-1}\mathbf{H}(\mathbf{N} - \mathbf{C})f_\mu + (\mathbf{I} + \mathbf{H}\mathbf{C})^{-1}e$$

gilt und f_0 ein beliebiges Element aus B^n ist. Weiter gilt für die μ -te Approximation f_μ von f

$$\|f_\mu - f\| \leq \frac{r(C, T)^\mu}{1 - r(C, T)} \|f_1 - f_0\|.$$

Beweis:

Zunächst wollen wir zeigen, daß die Operatorgleichung

$$f = -(I + HC)^{-1} H(N - C)f + (I + HC)^{-1} e \quad (5)$$

für alle $e \in B^n$ eine eindeutige Lösung $f \in B^n$ besitzt. Die Abbildung $\Phi: B^n \rightarrow B^n$ mit

$$h = \Phi f = -(I + HC)^{-1} H(N - C)f + (I + HC)^{-1} e$$

ist eine Kontraktion, da nach Abschnitt I.2.2 und nach den Voraussetzungen des Satzes für die inkrementale Verstärkung von

$$\begin{aligned} g_0(\Phi) &= g_0((I + HC)^{-1} H(N - C)) = g_0((I + HC)^{-1} HT^{-1} T(N - C)) \\ &\leq g_0((I + HC)^{-1} HT^{-1}) \cdot g_0(T(N - C)) \\ &\leq \|(I + HC)^{-1} HT^{-1}\| \cdot \eta(C, T) = r(C, T) < 1 \end{aligned}$$

gilt. Damit sind die Voraussetzungen des Banachschen Fixpunktsatzes erfüllt, und aus ihm folgen die Behauptungen bezüglich Gl. (5). Es muß jetzt noch gezeigt werden, daß eine Lösung $f \in B^n$ der Gl. (5) auch eine Lösung der Gl. (4) ist und umgekehrt:

Es sei mit $f \in B^n$ die Gleichung

$$f = -(I + HC)^{-1} H(N - C)f + (I + HC)^{-1} e$$

erfüllt. Dann ist

$$(I + HC)f = -H(N - C)f + e,$$

und daraus folgt sofort

$$e = f + \mathbf{H}\mathbf{N}f.$$

Umgekehrt sei $f \in \mathbf{B}^n$ eine Lösung der Gleichung

$$e = f + \mathbf{H}\mathbf{N}f.$$

Dann folgt

$$f + \mathbf{H}(\mathbf{N} - \mathbf{C})f + \mathbf{H}\mathbf{C}f = e,$$

$$(\mathbf{I} + \mathbf{H}\mathbf{C})f = -\mathbf{H}(\mathbf{N} - \mathbf{C})f + e$$

und schließlich

$$f = -(\mathbf{I} + \mathbf{H}\mathbf{C})^{-1}\mathbf{H}(\mathbf{N} - \mathbf{C})f + (\mathbf{I} + \mathbf{H}\mathbf{C})^{-1}e.$$

q.e.d.

Bemerkung 1:

Offensichtlich existiert der Operator $\mathbf{T}^{-1}$ genau dann, wenn $t_i \neq 0$ für $i=1, \dots, n$ gilt. Dann ist

$$\|\mathbf{T}^{-1}\| = \max_{i=1, \dots, n} 1/|t_i|.$$

Bemerkung 2:

Da nach Voraussetzung der Operator $(\mathbf{I} + \mathbf{H}\mathbf{C})$ beschränkt ist und $(\mathbf{I} + \mathbf{H}\mathbf{C})^{-1}$ existiert, ist nach Satz I.10 auch $(\mathbf{I} + \mathbf{H}\mathbf{C})^{-1}$ beschränkt. In den Voraussetzungen des Satzes 2 wurde deshalb nicht $\|(\mathbf{I} + \mathbf{H}\mathbf{C})^{-1}\mathbf{H}\mathbf{T}^{-1}\| < \infty$ gefordert, da dies eine Konsequenz ist. Das Analoge gilt für den folgenden Abschnitt.

Bemerkung 3:

Wie aus dem Beweis des Satzes 2 hervorgeht, stellt $r(\mathbf{C}, \mathbf{T})$ eine obere Schranke der Kontraktionskonstanten des Operators $(\mathbf{I} + \mathbf{H}\mathbf{C})^{-1}\mathbf{H}(\mathbf{N} - \mathbf{C})$ dar. Wir nennen deshalb $r(\mathbf{C}, \mathbf{T})$ Kontraktions-schranke.

2.4.2 Lösung der Gleichung $e = \mathbf{N}f + \mathbf{H}f$

Ein ähnlicher Beweis wie der des Satzes 2 führt unter den Voraussetzungen der Abschnitte II.2.1 und II.2.2 auf folgenden Satz:

Satz 3: Es seien die Operatoren $\mathbf{C}$ und $\mathbf{T}$ derart gegeben, daß folgende Bedingungen erfüllt sind:

- (i) $\mathbf{T}^{-1}$ existiert,
- (ii) $(\mathbf{C} + \mathbf{H})^{-1}$ existiert,
- (iii) $q(\mathbf{C}, \mathbf{T}) = \|(\mathbf{C} + \mathbf{H})^{-1} \mathbf{T}^{-1}\| \cdot \eta(\mathbf{C}, \mathbf{T}) < 1$.

Dann gibt es für jedes $e \in \mathbf{B}^n$ ein eindeutiges $f \in \mathbf{B}^n$, so daß

$$e = \mathbf{N}f + \mathbf{H}f$$

gilt. Ferner ist

$$f = \lim_{\mu \rightarrow \infty} f_{\mu},$$

wobei

$$f_{\mu+1} = -(\mathbf{C} + \mathbf{H})^{-1}(\mathbf{N} - \mathbf{C})f_{\mu} + (\mathbf{C} + \mathbf{H})^{-1}e$$

gilt und f_0 ein beliebiges Element von $\mathbf{B}^n$ ist.

Weiter gilt für die μ -te Approximation f_{μ} von f

$$\|f - f_{\mu}\| \leq \frac{q(\mathbf{C}, \mathbf{T})^{\mu}}{1 - q(\mathbf{C}, \mathbf{T})} \|f_1 - f_0\|.$$

Beweis:

Der Beweis und die Bemerkungen sind analog zum Abschnitt II.2.4.1.

2.4.3 Minimierung der Kontraktionsschranken

$r(\mathbf{C}, \mathbf{T})$ und $q(\mathbf{C}, \mathbf{T})$

Im folgenden soll der Einfluß der Transformationsoperatoren $\mathbf{C}$ und $\mathbf{T}$ auf die Kontraktionsschranken $r(\mathbf{C}, \mathbf{T})$ bzw. $q(\mathbf{C}, \mathbf{T})$ untersucht werden.

Es seien die Mengen

$$\Gamma_1 = \{ \mathbf{C} \mid (\mathbf{I} + \mathbf{H}\mathbf{C})^{-1} \text{ existiert} \},$$

$$\Gamma_2 = \{ \mathbf{C} \mid (\mathbf{C} + \mathbf{H})^{-1} \text{ existiert} \}$$

und

$$\Theta = \{ \mathbf{T} \mid \mathbf{T}^{-1} \text{ existiert} \}$$

gegeben.

Nun stellt sich die Frage, ob es ein "optimales" $\mathbf{T}_0^{(1)} \in \Theta$ und $\mathbf{C}_0^{(1)} \in \Gamma_1$ bzw. $\mathbf{T}_0^{(2)} \in \Theta$ und $\mathbf{C}_0^{(2)} \in \Gamma_2$ gibt, so daß die zugeordneten Kontraktionsschranken $r(\mathbf{C}_0^{(1)}, \mathbf{T}_0^{(1)})$ bzw. $q(\mathbf{C}_0^{(2)}, \mathbf{T}_0^{(2)})$ in einem gewissen Sinne minimal sind. In den folgenden Abschnitten wird gezeigt, daß solche optimale Transformationsoperatoren existieren.

Zunächst brauchen wir einige Definitionen:

Def. 2: Unter den Vektoren $\mathbf{c} = (c_1, c_2, \dots, c_n)^T \in \mathbf{R}^n$ sei der Vektor $\mathbf{c}_0 = (c_1^{(0)}, c_2^{(0)}, \dots, c_n^{(0)})^T$ durch

$$\mathbf{c}_0 = (\mathbf{a} + \mathbf{b})/2$$

definiert. Ferner sei unter den Vektoren

$\mathbf{t} = (t_1, t_2, \dots, t_n)^T \in \mathbf{R}^n$ der Vektor $\mathbf{t}(\mathbf{c}) = (t_1(c_1), t_2(c_2), \dots, t_n(c_n))^T$ durch

$$t_i(c_i) \eta_i(c_i) = t_j(c_j) \eta_j(c_j) \text{ für } i, j = 1, \dots, n$$

und $t_i(c_i) > 0$ für $i=1, \dots, n$ bis auf eine positive Konstante, die seine Länge bestimmt, gegeben. Diese Konstante sei willkürlich festgesetzt, da sie für das weitere ohne Bedeutung ist. Für $c = c_0$ setzen wir

$$t_0 = t(c_0) = (t_1^{(0)}, t_2^{(0)}, \dots, t_n^{(0)})^T.$$

Die Operatoren $\mathbf{C}, \mathbf{C}_0, \mathbf{T}, \mathbf{T}(\mathbf{C})$ und $\mathbf{T}_0$ sind im Sinne von Abschnitt II.2.2 durch die Vektoren $c, c_0, t, t(c)$ und t_0 definiert.

Bemerkung:

Wegen $\alpha_i < \beta_i$ für $i=1, \dots, n$ ist $\eta_i(c_i) > 0$ für $i=1, \dots, n$ und damit $t_i(c_i) > 0$ für $i=1, \dots, n$. Folglich ist $\mathbf{T}(\mathbf{C}) \in \Theta$ und insbesondere $\mathbf{T}_0 \in \Theta$.

2.4.3.1 Der Einfluß von α und β auf die Kontraktionsschranken

Die Kontraktionsschranken

$$r(\mathbf{C}, \mathbf{T}) = \|(I + \mathbf{H}\mathbf{C})^{-1} \mathbf{H}\mathbf{T}^{-1}\| \cdot \eta(\mathbf{C}, \mathbf{T})$$

bzw.

$$q(\mathbf{C}, \mathbf{T}) = \|(\mathbf{C} + \mathbf{H})^{-1} \mathbf{T}^{-1}\| \eta(\mathbf{C}, \mathbf{T})$$

hängen nur über $\eta(\mathbf{C}, \mathbf{T})$ von den Vektoren α und β ab. Minimale Kontraktionsschranken bedeuten also minimales $\eta(\mathbf{C}, \mathbf{T})$ und damit (siehe Bemerkung von Abschnitt II.2.2) maximale α_i und minimale β_i . Anders ausgedrückt: Um die Kontraktionsschranke möglichst klein zu machen, wählt man α_i und β_i so, daß bei Erfülltsein von Gl. (3) $\beta_i - \alpha_i$ für $i=1, \dots, n$ möglichst klein ist. Bei dieser Wahl von α_i und β_i stellt dann $\eta(\mathbf{C}, \mathbf{T})$ die beste Abschätzung für $g_0(\mathbf{T}(\mathbf{N}-\mathbf{C}))$ im Rahmen der Voraussetzungen dar. Wir werden also in den folgenden Abschnitten von vorgegebenen Vektoren α und β ausgehen und uns nur noch um den Einfluß der Transformationsoperatoren kümmern.

2.4.3.2 Minimierung von $r(\mathbf{C}, \mathbf{T})$

Zunächst soll gezeigt werden, daß die Länge des Vektors $\mathbf{t}$, der dem Transformationsoperator $\mathbf{T}$ zugeordnet ist, für die Kontraktionsschranke ohne Bedeutung ist und deshalb willkürlich gewählt werden darf (siehe Def. 2 von Abschnitt II.2.4.3).

Satz 4: Es sei $\mathbf{C} \in \Gamma$ und $\mathbf{T} \in \Theta$. Dann gilt für alle $\lambda \in \mathbf{R}$ mit $\lambda \neq 0$

- (i) $\lambda \mathbf{T} \in \Theta$,
- (ii) $r(\mathbf{C}, \mathbf{T}) = r(\mathbf{C}, \lambda \mathbf{T})$.

Beweis:

(i): Aus $t_i \neq 0$ und $\lambda \neq 0$ folgt $\lambda t_i \neq 0$ für $i=1, \dots, n$ und damit $\lambda \mathbf{T} \in \Theta$.

(ii): $r(\mathbf{C}, \lambda \mathbf{T})$

$$\begin{aligned}
 &= \|(\mathbf{I} + \mathbf{H}\mathbf{C})^{-1} \mathbf{H}(\lambda \mathbf{T})^{-1}\| \cdot \max_{i=1, \dots, n} (|\lambda \cdot t_i| \cdot \eta_i(c_i)) \\
 &= \|(\mathbf{I} + \mathbf{H}\mathbf{C})^{-1} \mathbf{H} \lambda^{-1} \mathbf{T}^{-1}\| \cdot |\lambda| \cdot \max_{i=1, \dots, n} (|t_i| \cdot \eta_i(c_i)) \\
 &= \|(\mathbf{I} + \mathbf{H}\mathbf{C})^{-1} \mathbf{H} \mathbf{T}^{-1}\| \cdot \max_{i=1, \dots, n} (|t_i| \cdot \eta_i(c_i)) \\
 &= r(\mathbf{C}, \mathbf{T}).
 \end{aligned}$$

q.e.d.

Imweiteren soll versucht werden, eine spezielle Wahl der Transformationsoperatoren $\mathbf{C} = \mathbf{C}_0$ und $\mathbf{T} = \mathbf{T}_0$ im Hinblick auf eine möglichst kleine Kontraktionsschranke zu rechtfertigen.

Wenn wir in Zukunft von den optimalen Transformationsoperatoren $\mathbf{C}_0$ und $\mathbf{T}_0$ sprechen, dann soll diese Optimalität im Sinne der folgenden Sätze 5 bis 11 zu verstehen sein.

Satz 5: Für einen vorgegebenen Transformationsoperator $\mathbf{C} \in \Gamma_1$ ist

$$r(\mathbf{C}, \mathbf{T}(\mathbf{C})) \leq r(\mathbf{C}, \mathbf{T})$$

für alle $\mathbf{T} \in \Theta$. Insbesondere gilt mit $\mathbf{C}_0 \in \Gamma_1$

$$r(\mathbf{C}_0, \mathbf{T}_0) \leq r(\mathbf{C}_0, \mathbf{T})$$

für alle $\mathbf{T} \in \Theta$.

Beweis:

$$r(\mathbf{C}, \mathbf{T}(\mathbf{C})) = \|(\mathbf{I} + \mathbf{H}\mathbf{C})^{-1} \mathbf{H} \mathbf{T}^{-1}(\mathbf{C})\| \cdot \eta(\mathbf{C}, \mathbf{T}(\mathbf{C}))$$

$$S1 \quad = \|(\mathbf{I} + \mathbf{H}\mathbf{C})^{-1} \mathbf{H} \mathbf{T}^{-1}(\mathbf{C}) \mathbf{T}^{-1} \mathbf{T}\| \cdot \eta(\mathbf{C}, \mathbf{T}(\mathbf{C}))$$

$$S2 \quad = \|(\mathbf{I} + \mathbf{H}\mathbf{C})^{-1} \mathbf{H} \mathbf{T}^{-1} \mathbf{T}(\mathbf{C})^{-1} \mathbf{T}\| \cdot \eta(\mathbf{C}, \mathbf{T}(\mathbf{C}))$$

$$S3 \quad \leq \|(\mathbf{I} + \mathbf{H}\mathbf{C})^{-1} \mathbf{H} \mathbf{T}^{-1}\| \cdot \|\mathbf{T}^{-1}(\mathbf{C}) \mathbf{T}\| \cdot \eta(\mathbf{C}, \mathbf{T}(\mathbf{C}))$$

$$S4 \quad = \|(\mathbf{I} + \mathbf{H}\mathbf{C})^{-1} \mathbf{H} \mathbf{T}^{-1}\| \cdot \max_i (|t_i| / t_i(c_i)) \cdot \max_i (t_i(c_i) \cdot \eta_i(c_i))$$

$$S5 \quad = \|(\mathbf{I} + \mathbf{H}\mathbf{C})^{-1} \mathbf{H} \mathbf{T}^{-1}\| \cdot \max_i \left(\frac{|t_i| \cdot \eta_i(c_i)}{t_i(c_i) \cdot \eta_i(c_i)} \right) \max_i (t_i(c_i) \cdot \eta_i(c_i))$$

$$S6 \quad = \|(\mathbf{I} + \mathbf{H}\mathbf{C})^{-1} \mathbf{H} \mathbf{T}^{-1}\| \cdot \max_i (|t_i| \cdot \eta_i(c_i))$$

$$S7 \quad = \|(\mathbf{I} + \mathbf{H}\mathbf{C})^{-1} \mathbf{H} \mathbf{T}^{-1}\| \cdot \eta(\mathbf{C}, \mathbf{T})$$

$$= r(\mathbf{C}, \mathbf{T}).$$

Den einzelnen Schritten liegen folgende Eigenschaften zugrunde:

S1 $\mathbf{T}^{-1}\mathbf{T} = \mathbf{I}$.

S2 $\mathbf{T}^{-1}(\mathbf{C})\mathbf{T}^{-1} = \mathbf{T}^{-1}\mathbf{T}^{-1}(\mathbf{C})$, da alle Transformationsoperatoren auf Grund ihrer Konstruktion untereinander kommutativ sind.

S3 $\|\mathbf{F}_1\mathbf{F}_2\| \leq \|\mathbf{F}_1\| \cdot \|\mathbf{F}_2\|$ (siehe Abschnitt I.2.2).

S4 Definition.

S5 Identische Erweiterung.

S6 Nach Konstruktion des Operators $\mathbf{T}(\mathbf{C})$ ist $t_i(c_i) \cdot \eta_i(c_i)$ eine von i unabhängige Konstante.

S7 Definition.

q.e.d.

Satz 6: Es existiere ein $\mathbf{C} \in \Gamma_1$ und $\mathbf{T} \in \Theta$ derart, daß

$$r(\mathbf{C}, \mathbf{T}) < 1$$

gilt. Dann ist $\mathbf{C}_0 \in \Gamma_1$.

Beweis:

Aus der Identität

$$\begin{aligned} (\mathbf{I} + \mathbf{H}\mathbf{C}_0) &\equiv (\mathbf{I} + \mathbf{H}\mathbf{C}) + \mathbf{H}(\mathbf{C}_0 - \mathbf{C}) \\ &\equiv (\mathbf{I} + \mathbf{H}\mathbf{C})(\mathbf{I} + (\mathbf{I} + \mathbf{H}\mathbf{C})^{-1}\mathbf{H}(\mathbf{C}_0 - \mathbf{C})) \\ &\equiv (\mathbf{I} + \mathbf{H}\mathbf{C})(\mathbf{I} + (\mathbf{I} + \mathbf{H}\mathbf{C})^{-1}\mathbf{H}\mathbf{T}^{-1}\mathbf{T}(\mathbf{C}_0 - \mathbf{C})) \end{aligned}$$

folgt mit Satz I.11, daß $(\mathbf{I} + \mathbf{H}\mathbf{C}_0)^{-1}$ existiert, falls

$$\|(\mathbf{I} + \mathbf{H}\mathbf{C})^{-1}\mathbf{H}\mathbf{T}^{-1}\mathbf{T}(\mathbf{C}_0 - \mathbf{C})\| < 1 \quad (6)$$

gilt.

Um nachzuweisen, daß Gl. (6) erfüllt ist, benötigen wir zunächst folgende Relationen, deren Gültigkeit anhand Abb. (3) eingesehen werden kann:

$$\eta_i(c_i) = \eta_i(c_i^{(o)}) + |c_i - c_i^{(o)}| \quad \text{für } i=1, \dots, n. \quad (7)$$

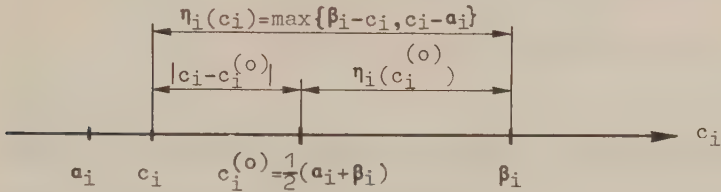


Abb. (3)

Aus Gl. (7) folgt mit $\eta_i(c_i^{(o)}) > 0$ für $i=1, \dots, n$

$$\eta_i(c_i) > |c_i - c_i^{(o)}| \quad \text{für } i=1, \dots, n,$$

und nach Multiplikation mit $|t_i|$ gilt

$$|t_i| \cdot \eta_i(c_i) > |t_i| \cdot |c_i - c_i^{(o)}| \quad \text{für } i=1, \dots, n.$$

Bildet man auf beiden Seiten das Maximum, so ergibt sich

$$\max_{i=1, \dots, n} (|t_i| \cdot \eta_i(c_i)) > \max_{i=1, \dots, n} (|t_i| \cdot |c_i - c_i^{(o)}|)$$

oder kurz

$$\eta(\mathbf{C}, \mathbf{T}) > \|\mathbf{T}(\mathbf{C}_0 - \mathbf{C})\|.$$

Damit ist

$$\begin{aligned} \|(I + \mathbf{H}\mathbf{C})^{-1} \mathbf{H} \mathbf{T}^{-1} \mathbf{T}(\mathbf{C}_0 - \mathbf{C})\| &\leq \|(I + \mathbf{H}\mathbf{C})^{-1} \mathbf{H} \mathbf{T}^{-1}\| \cdot \|\mathbf{T}(\mathbf{C}_0 - \mathbf{C})\| \\ &< \|(I + \mathbf{H}\mathbf{C})^{-1} \mathbf{H} \mathbf{T}^{-1}\| \cdot \eta(\mathbf{C}, \mathbf{T}) \\ &= r(\mathbf{C}, \mathbf{T}) < 1, \end{aligned}$$

und folglich existiert $(I + \mathbf{H}\mathbf{C}_0)^{-1}$.

q.e.d.

Satz 7: Es existiere ein $\mathbf{C} \in \Gamma_1$ derart, daß

$$r(\mathbf{C}, \mathbf{T}_0) < 1$$

gilt. Dann ist $\mathbf{C}_0 \in \Gamma_1$, und es gilt

$$r(\mathbf{C}_0, \mathbf{T}_0) \leq r(\mathbf{C}, \mathbf{T}_0)$$

für alle $\mathbf{C} \in \Gamma_1$.

Beweis:

$\mathbf{C}_0 \in \Gamma_1$ folgt aus Satz 6. Um die Ungleichung des Satzes zu beweisen, benötigen wir folgende Gleichungen (siehe Beweis von Satz 6):

$$\eta_i(c_i) = \eta_i(c_i^{(0)}) + |c_i - c_i^{(0)}| \quad \text{für } i=1, \dots, n.$$

Nach Multiplikation mit $t_i^{(0)}$ folgt

$$t_i^{(0)} \cdot \eta_i(c_i) = t_i^{(0)} \cdot \eta_i(c_i^{(0)}) + t_i^{(0)} \cdot |c_i - c_i^{(0)}| \quad \text{für } i=1, \dots, n.$$

Dabei ist $t_i^{(0)} \cdot \eta_i(c_i^{(0)})$ nach Konstruktion von $\mathbf{T}_0$ eine Konstante, die nicht von i abhängt, und somit ergibt sich

$$\max_{i=1, \dots, n} (t_i \cdot \eta_i(c_i)) = \max_{i=1, \dots, n} (t_i^{(0)} \cdot \eta_i(c_i^{(0)})) + \max_{i=1, \dots, n} (t_i^{(0)} \cdot |c_i - c_i^{(0)}|),$$

oder kurz

$$\eta(\mathbf{C}_0, \mathbf{T}_0) = \eta(\mathbf{C}, \mathbf{T}_0) - \|\mathbf{T}_0(\mathbf{C} - \mathbf{C}_0)\|.$$

Mit Hilfe dieser Gleichung gilt dann für alle $\mathbf{C} \in \Gamma_1$ unter Ausnützung der Dreiecksungleichung

$$\begin{aligned} r(\mathbf{C}, \mathbf{T}) &= \|(I + \mathbf{H}\mathbf{C}_0)^{-1} \mathbf{H} \mathbf{T}_0^{-1}\| \cdot \eta(\mathbf{C}_0, \mathbf{T}_0) \\ &= \|(I + \mathbf{H}\mathbf{C}_0)^{-1} \mathbf{H} \mathbf{T}_0^{-1}\| \cdot \eta(\mathbf{C}, \mathbf{T}_0) \\ &\quad - \|(I + \mathbf{H}\mathbf{C}_0)^{-1} \mathbf{H} \mathbf{T}_0^{-1}\| \cdot \|\mathbf{T}_0(\mathbf{C} - \mathbf{C}_0)\| \\ &\leq \|(I + \mathbf{H}\mathbf{C})^{-1} \mathbf{H} \mathbf{T}_0^{-1}\| \cdot \eta(\mathbf{C}, \mathbf{T}_0) \\ &\quad + \|(I + \mathbf{H}\mathbf{C}_0)^{-1} \mathbf{H} \mathbf{T}_0^{-1} - (I + \mathbf{H}\mathbf{C})^{-1} \mathbf{H} \mathbf{T}_0^{-1}\| \cdot \eta(\mathbf{C}, \mathbf{T}_0) \\ &\quad - \|(I + \mathbf{H}\mathbf{C}_0)^{-1} \mathbf{H} \mathbf{T}_0^{-1}\| \cdot \|\mathbf{T}_0(\mathbf{C} - \mathbf{C}_0)\|, \end{aligned}$$

und unter Verwendung der Abschätzung

$$\begin{aligned}
 & \| (I + HC_0)^{-1} H T_0^{-1} - (I + HC)^{-1} H T_0^{-1} \| \\
 &= \| ((I + HC_0)^{-1} - (I + HC)^{-1}) H T_0^{-1} \| \\
 &= \| (I + HC_0)^{-1} ((I + HC) - (I + HC_0)) (I + HC)^{-1} H T_0^{-1} \| \\
 &= \| (I + HC_0)^{-1} H T_0^{-1} T_0 (C - C_0) (I + HC)^{-1} H T_0^{-1} \| \\
 &\leq \| (I + HC_0)^{-1} H T_0^{-1} \| \cdot \| T_0 (C - C_0) \| \cdot \| (I + HC)^{-1} H T_0^{-1} \|
 \end{aligned}$$

folgt schließlich

$$r(C_0, T_0) \leq r(C, T_0) - (1 - r(C, T_0)) \cdot \| (I + HC_0)^{-1} H T_0^{-1} \| \cdot \| T_0 (C - C_0) \| ,$$

und damit die gewünschte Ungleichung.

q.e.d.

Die folgenden Sätze beantworten die Frage nach der Optimalität der Transformationsoperatoren C_0 und T_0 bezüglich aller $C \in \Gamma_1$ und $T \in \Theta$. Dabei werden wir grundsätzlich

$$T = T(C)$$

setzen, da diese Wahl nach Satz 5 für ein vorgegebenes $C \in \Gamma_1$ optimal ist. Zunächst benötigen wir folgenden Satz:

Satz 8: Es sei $C \in \Gamma_1$, und es gelte

$$r(C, T(C)) < 1$$

und

$$\min_{i=1, \dots, n} \left(\frac{|c_i - c_i^{(0)}|}{\beta_i - \alpha_i} \right) \geq r(C, T(C)) \cdot \max_{i=1, \dots, n} \left(\frac{|c_i - c_i^{(0)}|}{\beta_i - \alpha_i} \right) .$$

Dann ist $C_0 \in \Gamma_1$, und es gilt

$$r(C_0, T_0) \leq r(C, T(C)) .$$

Beweis:

$\mathbf{C}_0 \in \Gamma$ folgt aus Satz 6. Wie in den Beweisen der Sätze 6 und 7 gehen wir von den Gleichungen

$$\eta_i(c_i) = \eta_i(c_i^{(0)}) + |c_i - c_i^{(0)}| \quad \text{für } i=1, \dots, n \quad (8)$$

aus.

Nach Multiplikation mit $t_i^{(0)}$ und Maximumbildung auf beiden Seiten folgt

$$\eta(\mathbf{C}, \mathbf{T}_0) = \eta(\mathbf{C}_0, \mathbf{T}_0) + \|\mathbf{T}_0(\mathbf{C} - \mathbf{C}_0)\| \quad (9)$$

Nach Definition von $\mathbf{T}_0$ und $\mathbf{T}(\mathbf{C})$ ist

$$t_i(c_i) \cdot \eta_i(c_i) = t_j(c_j) \cdot \eta_j(c_j) \quad \text{für } i, j=1, \dots, n \quad (10)$$

$$t_i^{(0)} \cdot \eta_i(c_i^{(0)}) = t_j^{(0)} \cdot \eta_j(c_j^{(0)}) \quad \text{für } i, j=1, \dots, n. \quad (11)$$

Setzt man Gl. 8 in Gl. 11 ein, so folgt

$$t_i^{(0)} \cdot (\eta_i(c_i) - |c_i - c_i^{(0)}|) = t_j^{(0)} (\eta_j(c_j) - |c_j - c_j^{(0)}|) \quad \text{für } i, j=1, \dots, n$$

bzw.

$$\underbrace{t_i^{(0)} \left(1 - \frac{|c_i - c_i^{(0)}|}{\eta_i(c_i)} \right)}_{t_i(c_i)} = \underbrace{t_j^{(0)} \left(1 - \frac{|c_j - c_j^{(0)}|}{\eta_j(c_j)} \right)}_{t_j(c_j)} \eta_j(c_j) \quad \text{für } i, j=1, \dots, n.$$

Die Klammern deuten die gemäß Gl. (10) durchgeführte Identifikation an. Diese Identifikation kann immer durchgeführt werden, indem man den Vektor $t(c)$ geeignet normiert, was auf die Kontraktionsschranke $r(\mathbf{C}, \mathbf{T}(\mathbf{C}))$ nach Satz 4 keinen Einfluß hat. Durch diese Normierung sind die Längen der Vektoren $t(c)$ und t_0 miteinander gekoppelt. Es gilt also

$$t_i(c_i) = t_i^{(0)} \left(1 - \frac{|c_i - c_i^{(0)}|}{\eta(c_i)} \right) \quad \text{für } i=1, \dots, n \quad (12)$$

bzw.

$$t_i(c_i) \eta_i(c_i) = t_i^{(0)} \cdot \eta_i(c_i) - t_i^{(0)} |c_i - c_i^{(0)}| \quad \text{für } i=1, \dots, n.$$

Nach einer Maximumbildung auf beiden Seiten folgt

$$\eta(\mathbf{C}, \mathbf{T}_0) = \eta(\mathbf{C}, \mathbf{T}(\mathbf{C})) + \|\mathbf{T}_0(\mathbf{C} - \mathbf{C}_0)\|. \quad (13)$$

Mit den Gl. (9) und (13) gilt schließlich

$$\eta(\mathbf{C}_0, \mathbf{T}_0) = \eta(\mathbf{C}, \mathbf{T}(\mathbf{C})).$$

Um die Ungleichung des Satzes zu beweisen, genügt es also zu zeigen, daß

$$\|(\mathbf{I} + \mathbf{H}\mathbf{C}_0)^{-1} \mathbf{H}\mathbf{T}_0^{-1}\| \leq \|(\mathbf{I} + \mathbf{H}\mathbf{C})^{-1} \mathbf{H}\mathbf{T}^{-1}(\mathbf{C})\| \quad (14)$$

ist. (Das war der Grund für die Normierung von $t(c)$.)

Mit den Abkürzungen $\mathbf{T} = \mathbf{T}(\mathbf{C})$,

$$\mathbf{S}_0 = (\mathbf{I} + \mathbf{H}\mathbf{C}_0)^{-1} \mathbf{H}\mathbf{T}_0^{-1}$$

und

$$\mathbf{S} = (\mathbf{I} + \mathbf{H}\mathbf{C})^{-1} \mathbf{H}\mathbf{T}^{-1}$$

muß also nachgewiesen werden, daß

$$\|\mathbf{S}_0\| \leq \|\mathbf{S}\| \quad (15)$$

ist.

Es gilt (siehe Umformung im Beweis von Satz 7)

$$\begin{aligned} \mathbf{S}_0 &= (\mathbf{I} + \mathbf{H}\mathbf{C})^{-1} \mathbf{H}\mathbf{T}_0^{-1} + ((\mathbf{I} + \mathbf{H}\mathbf{C}_0)^{-1} - (\mathbf{I} + \mathbf{H}\mathbf{C})^{-1}) \mathbf{H}\mathbf{T}_0^{-1} \\ &= \mathbf{S}\mathbf{T}_0^{-1}\mathbf{T} + (\mathbf{I} + \mathbf{H}\mathbf{C}_0)^{-1} \mathbf{H}(\mathbf{C} - \mathbf{C}_0)(\mathbf{I} + \mathbf{H}\mathbf{C})^{-1} \mathbf{H}\mathbf{T}_0^{-1} \\ &= \mathbf{S}\mathbf{T}_0^{-1}\mathbf{T} + \mathbf{S}_0\mathbf{T}_0(\mathbf{C} - \mathbf{C}_0)\mathbf{S}\mathbf{T}_0^{-1}\mathbf{T} \end{aligned}$$

oder

$$S_0(I - T_0(C - C_0)S T_0^{-1}T) = S T_0^{-1}T$$

$$S_0(T_0 T^{-1} - T_0(C - C_0)S) = S$$

$$S_0 T_0 T^{-1}(I - T(C - C_0)S) = S.$$

Wegen $\|T(C - C_0)S\| < 1$ existiert $(I - T(C - C_0)S)^{-1}$, und es folgt

$$S_0 = S(I - T(C - C_0)S)^{-1} T T_0^{-1}$$

oder

$$S_0 = S(T_0 T^{-1} - T_0(C - C_0)S)^{-1}. \quad (16)$$

Es ist interessant, daß die Kopplung zwischen den Operatoren S_0 und S in dieser einfachen Form angegeben werden kann.

Aus Gl. (16) folgt

$$\|S_0\| \leq \|S\| \cdot \|(T_0 T^{-1} - T_0(C - C_0)S)^{-1}\|.$$

Um Gl. (15) zu beweisen, genügt es also zu zeigen, daß

$$\|(T_0 T^{-1} - T_0(C - C_0)S)^{-1}\| \leq 1$$

gilt. Für alle $h \in B^n$ ist

$$\|(T_0 T^{-1} - T_0(C - C_0)S)h\| \geq \|T_0 T^{-1}h\| - \|T_0(C - C_0)Sh\|. \quad (17)$$

Nach Gl. (12) gilt

$$\frac{t_i^{(0)}}{t_i(c_i)} = \frac{1}{1 - \frac{|c_i - c_i^{(0)}|}{\eta_i(c_i)}} = 1 + \frac{|c_i - c_i^{(0)}|}{\eta_i(c_i^{(0)})} \quad \text{für } i=1, \dots, n$$

und damit

$$\|T_0 T^{-1} h\| \geq \min_{i=1, \dots, n} \left(\frac{t_i^{(0)}}{t_i(c_i)} \right) \cdot \|h\| = \left(1 + \min_{i=1, \dots, n} \left(\frac{|c_i - c_i^{(0)}|}{\eta_i(c_i^{(0)})} \right) \right) \cdot \|h\| \quad (18)$$

für alle $h \in B^n$ (siehe (iii) von Abschnitt II.2.1).

Ferner ist

$$\begin{aligned} \|T_0(C - C_0)Sh\| &\leq \|T_0(C - C_0)\| \cdot \|S\| \cdot \|h\| \\ &= \max_{i=1, \dots, n} (t_i^{(0)} \cdot |c_i - c_i^{(0)}|) \cdot \|S\| \cdot \|h\| \\ &= \max_{i=1, \dots, n} (t_i^{(0)} \cdot \eta_i(c_i^{(0)}) \cdot \frac{|c_i - c_i^{(0)}|}{\eta_i(c_i^{(0)})}) \cdot \|S\| \cdot \|h\| \\ &= \max_{i=1, \dots, n} \left(\frac{|c_i - c_i^{(0)}|}{\eta_i(c_i^{(0)})} \right) \cdot r(C, T) \cdot \|h\| \end{aligned} \quad (19)$$

für alle $h \in B^n$.

Aus Gl. (17), (18) und (19) folgt

$$\begin{aligned} \|(T_0 T^{-1} - T_0(C - C_0)S)h\| &\geq \\ &\left\{ 1 + \min_{i=1, \dots, n} \left(\frac{|c_i - c_i^{(0)}|}{\eta_i(c_i^{(0)})} \right) - r(C, T) \cdot \max_{i=1, \dots, n} \left(\frac{|c_i - c_i^{(0)}|}{\eta_i(c_i^{(0)})} \right) \right\} \cdot \|h\| \end{aligned}$$

für alle $h \in B^n$.

Mit der getroffenen Voraussetzung

$$\min_{i=1, \dots, n} \left(\frac{|c_i - c_i^{(0)}|}{\eta_i(c_i^{(0)})} \right) \geq r(C, T) \cdot \max_{i=1, \dots, n} \left(\frac{|c_i - c_i^{(0)}|}{\eta_i(c_i^{(0)})} \right)$$

erhalten wir

$$\|(T_0 T^{-1} - T_0(C - C_0)S)h\| \geq \|h\|$$

für alle $h \in B^n$ und nach Satz I.9 folgt schließlich Gl. (15).

q.e.d.

Satz 9: Es sei $\mathbf{C} \in \Gamma_1$, und es gelte

$$r(\mathbf{C}, \mathbf{T}(\mathbf{C})) < 1. \quad (20)$$

Unter diesen Voraussetzungen ist es nicht möglich zu beweisen, daß

$$r(\mathbf{C}_0, \mathbf{T}_0) \leq r(\mathbf{C}, \mathbf{T}(\mathbf{C}))$$

gilt.

Beweis:

Der Beweis erfolgt anhand eines Gegenbeispiels, für das unter den gegebenen Voraussetzungen

$$r(\mathbf{C}_0, \mathbf{T}_0) > r(\mathbf{C}, \mathbf{T}(\mathbf{C}))$$

gilt. Bei der Konstruktion dieses Gegenbeispiels knüpfen wir an den Beweis von Satz 8 an und verwenden die dort vereinbarten Abkürzungen. Es war

$$\mathbf{S}_0 = \mathbf{S}(\mathbf{T}_0 \mathbf{T}^{-1} - \mathbf{T}_0(\mathbf{C} - \mathbf{C}_0)\mathbf{S})^{-1}$$

bzw.

$$\mathbf{S}_0(\mathbf{T}_0 \mathbf{T}^{-1} - \mathbf{T}_0(\mathbf{C} - \mathbf{C}_0)\mathbf{S}) = \mathbf{S}. \quad (21)$$

Wir setzen nun

$$\mathbf{T}_0 \mathbf{T}^{-1} = \mathbf{I} + \mathbf{R}$$

mit $(\mathbf{R}h)_i = r_i \cdot h_i$ für $i=1, \dots, n$ und $h \in \mathbf{B}^n$. Es ist also

$$r_i = \frac{|c_i - c_i^{(0)}|}{\eta_i(c_i^{(0)})} \quad \text{für } i=1, \dots, n. \quad (22)$$

Ferner sei

$$-\mathbf{T}_0(\mathbf{C}_0 - \mathbf{C}) = \mathbf{R}\mathbf{A}$$

mit $(\Lambda h)_i = \lambda_i \cdot h_i$ für $i=1, \dots, n$, wobei $h \in \mathbf{B}^n$ und

$$\lambda_i = t_i(c_i^{(0)}) \cdot \eta_i(c_i^{(0)}) \cdot \text{sign}(c_i - c_i^{(0)}) \quad (23)$$

ist. Jetzt wählen wir

$$t_i(c_i^{(0)}) \cdot \eta_i(c_i^{(0)}) = t_i(c_i) \cdot \eta_i(c_i) = 1 \quad \text{für } i=1, \dots, n. \quad (24)$$

Damit ist

$$\mathbf{T}_0 \mathbf{T}^{-1} - \mathbf{T}_0 (\mathbf{C} - \mathbf{C}_0) \mathbf{S} = \mathbf{I} + \mathbf{R} (\mathbf{I} - \Lambda \mathbf{S})$$

mit $\|\Lambda\| = 1$ und $\|\Lambda \mathbf{S}\| \leq \|\Lambda\| \cdot \|\mathbf{S}\| = r(\mathbf{C}, \mathbf{T}) < 1$.

Für die hier durchgeführte Normierung der Transformationsoperatoren $\mathbf{T}$ und $\mathbf{T}_0$ ist also

$$\|\mathbf{S}\| < 1. \quad (25)$$

Mit den gewählten Bezeichnungen geht Gl. (21) über in

$$\mathbf{S}_0 (\mathbf{I} + \mathbf{R} (\mathbf{I} - \Lambda \mathbf{S})) = \mathbf{S}. \quad (26)$$

Gemäß Gl. (20) existiert $\mathbf{S}_0$, falls mit den nach Gl. (24) konstruierten Operatoren $\mathbf{T}_0$ und $\mathbf{T}$ die Gl. (25) gilt (siehe Satz 6).

Für unser Gegenbeispiel wählen wir nun als Banachraum $\mathbf{B}^n$ den reellen euklidischen Raum $\mathbf{R}^2$ mit der Norm

$$\|h\| = \sqrt{h_1^2 + h_2^2} \quad \text{für } h = (h_1, h_2)^T \in \mathbf{R}^2.$$

Sämtliche Operatoren sind dann reelle 2×2 -Matrizen. Die Diagonalmatrix $\mathbf{R}$ ist nach Gl. (22) positiv semidefinit. Wir wollen annehmen, daß die Diagonalelemente der Diagonalmatrix Λ alle positiv sind, was nach Gl. (23) immer durch die Wahl

$$c_i > c_i^{(0)} \quad \text{für } i=1, 2$$

zu erreichen ist. Damit gilt

$$\Lambda = 1_2 \quad (1_2 = 2 \times 2\text{-Einheitsmatrix}),$$

und Gl. (26) wird zu

$$S_0 (1_2 + R(1_2 - S)) = S.$$

Nach diesen Vereinbarungen wollen wir die Konstruktionsidee anhand des Vektordiagramms von Abb. (4) erläutern, da hier die Norm eines Vektors direkt seiner Länge entspricht. Es sei

$$M = 1_2 + R(1_2 - S),$$

d.h.

$$S_0 M = S$$

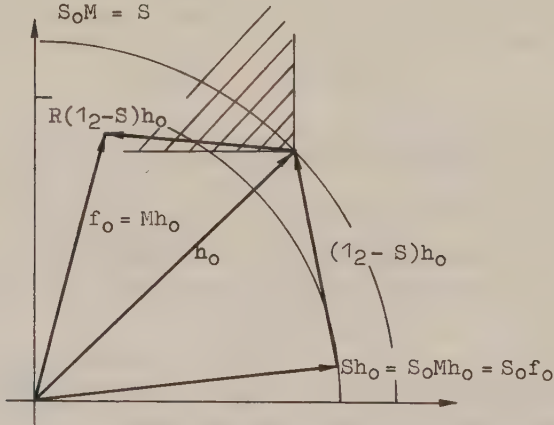


Abb. (4)

Es sei $h_0 \in \mathbb{R}^2$ der Vektor, den die Matrix S am meisten streckt. Wegen

$$\|Sh_0\| = \|S\| \cdot \|h_0\| < \|h_0\|$$

(siehe Abb. (4)) gilt wie gefordert

$$\|S\| < 1.$$

Die Vektoren h_0 und Sh_0 seien ferner wie in Abb. (4) ersichtlich durch eine geeignete Wahl der Matrix S festgelegt. Da die Matrix R positiv semidefinit ist, bildet R den Vektor $(1_2 - S)h_0$ prinzipiell in das schraffierte Gebiet ab. Wir wählen nun R nach Abb. (4) so, daß

$$\|S_0 f_0\| > \|f_0\| \quad \text{mit } f_0 = Mh_0$$

gilt. Folglich ist $\|S_0\| > 1$, und damit gilt $\|S\| < \|S_0\|$.

Wegen

$$\eta(C_0, T_0) = \eta(C, T(C))$$

ist schließlich

$$r(C_0, T_0) > r(C, T(C)),$$

womit unser Satz bewiesen ist.

Der Vollständigkeit halber wollen wir jetzt noch ein Gegenbeispiel, dem das geschilderte Konstruktionsverfahren zugrunde liegt, explizit angeben. Dabei seien folgende Größen vorgegeben:

$$a_1 = -1 \quad \beta_1 = 0 \quad c_1 = 0$$

$$a_2 = -1 \quad \beta_2 = 1 \quad c_2 = 0$$

$$H = \begin{pmatrix} 0,9 & 0,4 \\ 0 & 0 \end{pmatrix}.$$

Daraus folgt

$$c_1^{(0)} = -0,5 \quad \eta_1(c_1^{(0)}) = 0,5 \quad t_1^{(0)} = 2 \quad \eta_1(c_1) = 1 \quad t_1(c_1) = 1$$

$$c_2^{(0)} = 0 \quad \eta_2(c_2^{(0)}) = 1 \quad t_2^{(0)} = 1 \quad \eta_2(c_2) = 1 \quad t_2(c_2) = 1$$

und

$$S = (1_n + HC)^{-1} H T^{-1} = H T^{-1} = H$$

mit der Norm

$$\|S\| = \sqrt{\lambda_{\max}(S^T S)} = \sqrt{0,97} < 1.$$

Weiter ist

$$S_0 = (I_n + H C_0)^{-1} H T_0^{-1}$$

$$= \left(\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} 0,9 & 0,4 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} -0,5 & 0 \\ 0 & 0 \end{pmatrix} \right)^{-1} \begin{pmatrix} 0,9 & 0,4 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 0,5 & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0,9/1,1 & 0,8/1,1 \\ 0 & 0 \end{pmatrix}$$

mit der Norm

$$\|S_0\| = \sqrt{\lambda_{\max}(S_0^T S_0)} = \sqrt{\frac{1,45}{1,21}} > 1.$$

Also gilt

$$r(C_0, T_0) = \sqrt{0,97} < 1,$$

$$r(C, T(C)) = \sqrt{\frac{1,45}{1,21}} > 1$$

und damit

$$r(C_0, T_0) > r(C, T(C)).$$

q.e.d.

Im weiteren benötigen wir folgende Definition:

Def. 3: Für jeden Transformationsoperator $\hat{C} \in \Gamma_1$ ist die Klasse $K_1(\hat{C})$ von Transformationsoperatoren C durch

$$K_1(\hat{C}) = \left\{ C \in \Gamma_1 \mid \min_{i=1, \dots, n} \left(\frac{|c_i - c_i^{(0)}|}{\beta_i - \alpha_i} \right) \right. \\ \left. \geq r(\hat{C}, T(\hat{C})) \cdot \max_{i=1, \dots, n} \left(\frac{|c_i - c_i^{(0)}|}{\beta_i - \alpha_i} \right) \right\}$$

definiert.

Bemerkung:

Für $r(\hat{\mathbf{C}}, \mathbf{T}(\hat{\mathbf{C}})) > 1$ ist die Menge $\mathbf{K}_1(\hat{\mathbf{C}})$ gleich $\{\mathbf{C}_0\} \cap \Gamma_1$.

Satz 10: Es sei $\hat{\mathbf{C}} \in \mathbf{K}_1(\hat{\mathbf{C}})$, und es gelte

$$r(\hat{\mathbf{C}}, \mathbf{T}(\hat{\mathbf{C}})) < 1.$$

Dann ist $\mathbf{C}_0 \in \mathbf{K}_1(\hat{\mathbf{C}})$, und es gilt

$$r(\mathbf{C}_0, \mathbf{T}_0) \leq r(\mathbf{C}, \mathbf{T}(\mathbf{C}))$$

für alle $\mathbf{C} \in \mathbf{K}_1(\hat{\mathbf{C}})$.

Beweis:

Mit $\mathbf{C} \in \mathbf{K}_1(\hat{\mathbf{C}})$ sei

$$r(\mathbf{C}, \mathbf{T}(\mathbf{C})) \geq r(\hat{\mathbf{C}}, \mathbf{T}(\hat{\mathbf{C}})).$$

Damit ist die zu beweisende Ungleichung sicher erfüllt, da nach Satz 8

$$r(\mathbf{C}_0, \mathbf{T}_0) \leq r(\hat{\mathbf{C}}, \mathbf{T}(\hat{\mathbf{C}}))$$

gilt. Für

$$r(\mathbf{C}, \mathbf{T}(\mathbf{C})) < r(\hat{\mathbf{C}}, \mathbf{T}(\hat{\mathbf{C}}))$$

sind die Voraussetzungen des Satzes 8 erfüllt, da aus

$$r(\mathbf{C}, \mathbf{T}(\mathbf{C})) < r(\hat{\mathbf{C}}, \mathbf{T}(\hat{\mathbf{C}}))$$

sofort $\mathbf{K}_1(\hat{\mathbf{C}}) \subset \mathbf{K}_1(\mathbf{C})$ und damit $\mathbf{C} \in \mathbf{K}_1(\mathbf{C})$ folgt.

q.e.d.

Nach Satz 10 ist also der Transformationsoperator $\mathbf{C}_0$ bezüglich einer bestimmten Klasse von Operatoren optimal, falls in dieser Klasse überhaupt irgendein Operator $\hat{\mathbf{C}}$ zu finden ist, für den

$$r(\hat{\mathbf{C}}, \mathbf{T}(\hat{\mathbf{C}})) < 1$$

gilt. Anders ausgedrückt, falls

$$r(\mathbf{C}_0, \mathbf{T}_0) \geq 1$$

ist, gibt es sicher kein $\hat{\mathbf{C}} \in \mathbf{K}_1(\hat{\mathbf{C}})$ mit

$$r(\hat{\mathbf{C}}, \mathbf{T}(\hat{\mathbf{C}})) < 1.$$

Wie man anhand der Definition leicht ersieht, ist die Klasse $\mathbf{K}_1(\mathbf{C})$ um so allgemeiner, je kleiner $r(\mathbf{C}, \mathbf{T}(\mathbf{C}))$ ist. Falls irgendein $\hat{\mathbf{C}} \in \mathbf{K}_1(\hat{\mathbf{C}})$ mit $r(\hat{\mathbf{C}}, \mathbf{T}(\hat{\mathbf{C}})) < 1$ existiert, gilt

$$r(\mathbf{C}_0, \mathbf{T}_0) \leq r(\hat{\mathbf{C}}, \mathbf{T}(\hat{\mathbf{C}})) < 1$$

und somit

$$\mathbf{K}_1(\hat{\mathbf{C}}) \subseteq \mathbf{K}_1(\mathbf{C}_0).$$

Die allgemeinste Klasse, bezüglich der $\mathbf{C}_0$ optimal ist (falls überhaupt eine existiert), ist also $\mathbf{K}_1(\mathbf{C}_0)$. Der folgende Spezialfall von Satz 10 ist daher am aussagekräftigsten:

Satz 11: Es sei $r(\mathbf{C}_0, \mathbf{T}_0) < 1$ mit $\mathbf{C}_0 \in \mathbf{F}_1$. Dann ist

$$r(\mathbf{C}_0, \mathbf{T}_0) \leq r(\mathbf{C}, \mathbf{T}(\mathbf{C}))$$

für alle $\mathbf{C} \in \mathbf{K}_1(\mathbf{C}_0)$.

Beweis:

Trivialerweise gilt $\mathbf{C}_0 \in \mathbf{K}_1(\mathbf{C}_0)$. Damit sind die Voraussetzungen von Satz 10 erfüllt.

q.e.d.

Im Falle $n=2$ kann die Optimalitätsklasse $K_1(C_0)$ leicht veranschaulicht werden. Jeder Operator C kann hier in einer c_1, c_2 -Ebene durch einen Punkt definiert werden. In den folgenden Abbildungen sind in verschiedenen Fällen die Optimalitätsklassen $K_1(C_0)$ in die c_1, c_2 -Ebene schraffiert eingezeichnet.

Fall 1:

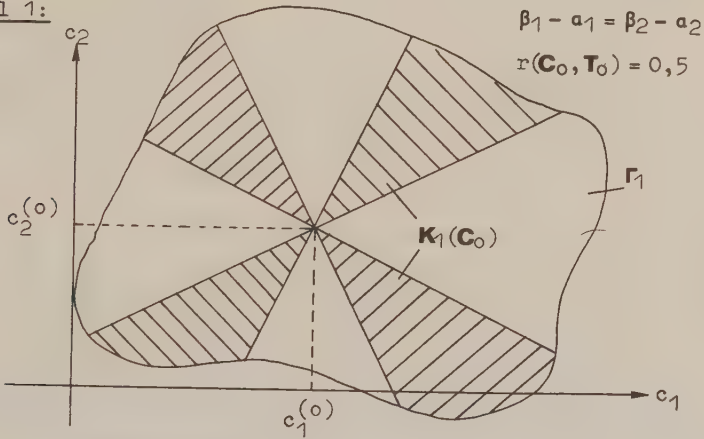


Abb. (5)

Fall 2:

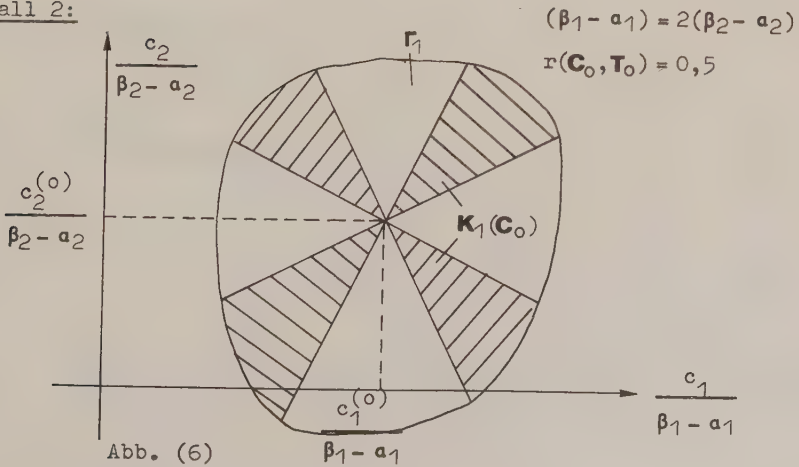


Abb. (6)

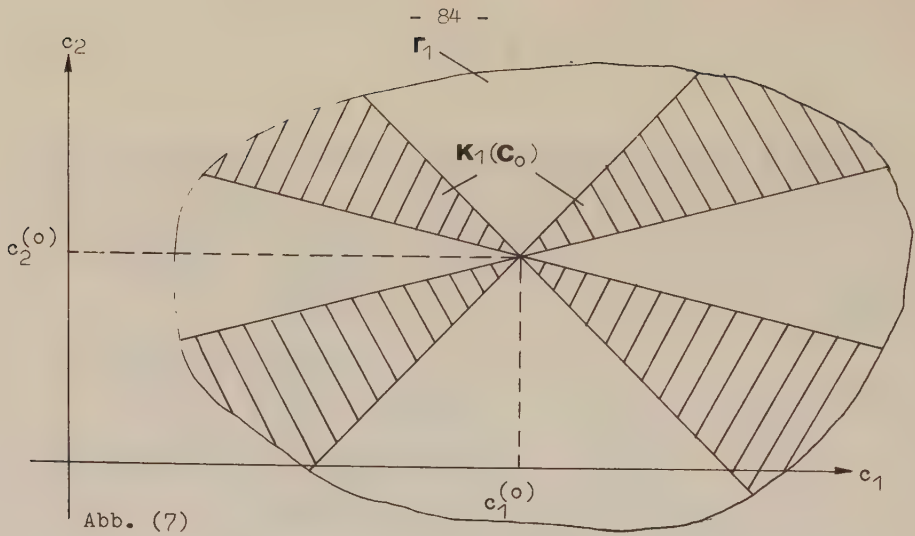


Abb. (7)

Fall 3:

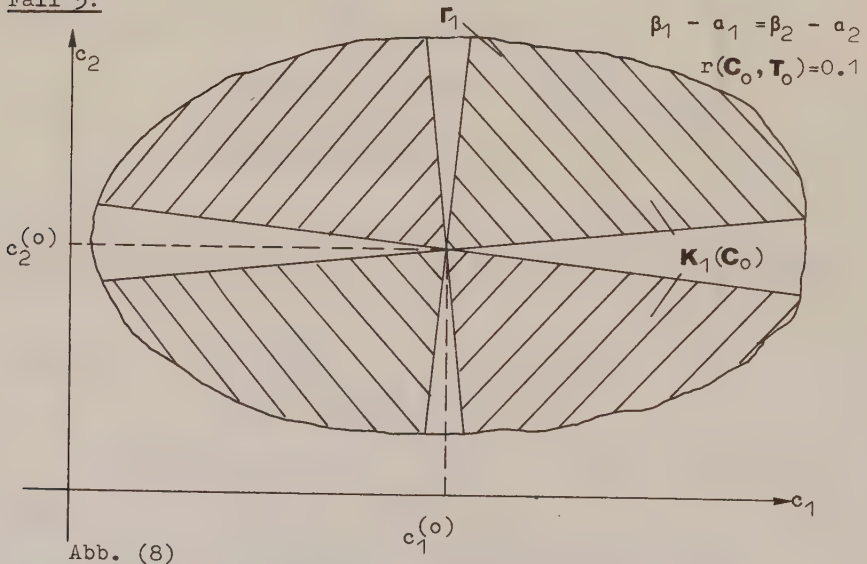


Abb. (8)

Wie man aus den Abb. (5) bis (8) entnehmen kann, "strebt" $K_1(C_0)$ für $r(C_0, T_0) \rightarrow 0$ gegen Γ_1 .

2.4.3.3 Minimierung von $q(\mathbf{C}, \mathbf{T})$

Die folgenden Sätze sind zu denen des vorhergehenden Abschnitts völlig analog und werden ohne Beweis angegeben.

Satz 12: Es sei $\mathbf{C} \in \Gamma_2$ und $\mathbf{T} \in \Theta$. Dann gilt für alle $\lambda \in \mathbf{R}$ mit $\lambda \neq 0$

$$(i) \quad \lambda \mathbf{T} \in \Theta,$$

$$(ii) \quad q(\mathbf{C}, \mathbf{T}) = q(\mathbf{C}, \lambda \mathbf{T})$$

Satz 13: Für einen vorgegebenen Transformationsoperator $\mathbf{C} \in \Gamma_2$ gilt

$$q(\mathbf{C}, \mathbf{T}(\mathbf{C})) \leq q(\mathbf{C}, \mathbf{T})$$

für alle $\mathbf{T} \in \Theta$. Insbesondere ist mit $\mathbf{C}_0 \in \Gamma_2$

$$q(\mathbf{C}_0, \mathbf{T}_0) \leq q(\mathbf{C}_0, \mathbf{T})$$

für alle $\mathbf{T} \in \Theta$.

Satz 14: Es existiere ein $\mathbf{C} \in \Gamma_2$ und $\mathbf{T} \in \Theta$ derart, daß

$$q(\mathbf{C}, \mathbf{T}) < 1$$

gilt. Dann ist

$$\mathbf{C}_0 \in \Gamma_2.$$

Satz 15: Es existiere ein $\mathbf{C} \in \Gamma_2$ derart, daß

$$q(\mathbf{C}, \mathbf{T}_0) < 1$$

gilt. Dann ist $\mathbf{C}_0 \in \Gamma_2$, und es gilt

$$q(\mathbf{C}_0, \mathbf{T}_0) \leq q(\mathbf{C}, \mathbf{T}_0)$$

für alle $\mathbf{C} \in \Gamma_2$.

Satz 16: Es sei $\mathbf{C} \in \Gamma_2$, und es gelte

$$q(\mathbf{C}, \mathbf{T}(\mathbf{C})) < 1$$

und

$$\min_{i=1, \dots, n} \left(\frac{|c_i - c_i^{(0)}|}{\beta_i - \alpha_i} \right) \geq q(\mathbf{C}, \mathbf{T}(\mathbf{C})) \cdot \max_{i=1, \dots, n} \left(\frac{|c_i - c_i^{(0)}|}{\beta_i - \alpha_i} \right).$$

Dann ist $\mathbf{C}_0 \in \Gamma_2$, und es gilt

$$q(\mathbf{C}_0, \mathbf{T}_0) \leq q(\mathbf{C}, \mathbf{T}(\mathbf{C})).$$

Satz 17: Es sei $\mathbf{C} \in \Gamma_2$, und es gelte

$$q(\mathbf{C}, \mathbf{T}(\mathbf{C})) < 1.$$

Unter diesen Voraussetzungen ist es nicht möglich zu beweisen, daß

$$q(\mathbf{C}_0, \mathbf{T}_0) \leq q(\mathbf{C}, \mathbf{T}(\mathbf{C}))$$

gilt.

Def. 4: Für jeden Transformationsoperator $\hat{\mathbf{C}} \in \Gamma_2$ ist die Klasse $\mathbf{K}_2(\hat{\mathbf{C}})$ von Transformationsoperatoren $\mathbf{C}$ durch

$$\begin{aligned} \mathbf{K}_2(\hat{\mathbf{C}}) &= \left\{ \mathbf{C} \in \Gamma_2 \mid \min_{i=1, \dots, n} \left(\frac{|c_i - c_i^{(0)}|}{\beta_i - \alpha_i} \right) \right. \\ &\quad \left. \geq q(\hat{\mathbf{C}}, \mathbf{T}(\hat{\mathbf{C}})) \cdot \max_{i=1, \dots, n} \left(\frac{|c_i - c_i^{(0)}|}{\beta_i - \alpha_i} \right) \right\} \end{aligned}$$

definiert.

Bemerkung:

Für $q(\hat{\mathbf{C}}, \mathbf{T}(\hat{\mathbf{C}})) > 1$ ist die Menge $\mathbf{K}_2(\hat{\mathbf{C}})$ gleich $\{\mathbf{C}_0\} \cap \Gamma_2$.

Satz 18: Es sei $\hat{\mathbf{C}} \in \mathbf{K}_2(\hat{\mathbf{C}})$, und es gelte

$$q(\hat{\mathbf{C}}, \mathbf{T}(\hat{\mathbf{C}})) < 1.$$

Dann ist $\mathbf{C}_0 \in \mathbf{K}_2(\hat{\mathbf{C}})$, und es gilt

$$q(\mathbf{C}_0, \mathbf{T}_0) \leq q(\mathbf{C}, \mathbf{T}(\mathbf{C}))$$

für alle $\mathbf{C} \in \mathbf{K}_2(\hat{\mathbf{C}})$.

Satz 19: Es sei $q(\mathbf{C}_0, \mathbf{T}_0) < 1$ mit $\mathbf{C}_0 \in \mathbf{F}_2$. Dann ist

$$q(\mathbf{C}_0, \mathbf{T}_0) \leq q(\mathbf{C}, \mathbf{T}(\mathbf{C}))$$

für alle $\mathbf{C} \in \mathbf{K}_2(\mathbf{C}_0)$.

2.4.3.4 Diskussion und Vergleich mit den Ergebnissen von Sandberg

I.W. Sandberg hat in [15] die Lösbarkeit der Operatorgleichungen (1) und (2) unter den speziellen Voraussetzungen

$$\alpha_i = \alpha_j = \tilde{\alpha}, \quad \beta_i = \beta_j = \tilde{\beta} \quad \text{für } i, j = 1, \dots, n \quad (27)$$

eingehend diskutiert. Der folgende Vergleich der Ergebnisse bezieht sich auf Gl. (1) und gilt sinngemäß auch für Gl. (2).

Sandberg beschränkt sich bei seiner Untersuchung auf einen Transformationsoperator $\mathbf{C}$ der Form $\mathbf{C} = \gamma \mathbf{I}$ mit $\gamma \in \mathbf{R}$. Der Einfluß des Transformationsoperators $\mathbf{T}$ wird von Sandberg nicht diskutiert. (Bei ihm ist $\mathbf{T} = \mathbf{I}$.)

Mit der hier vereinbarten Terminologie hat Sandberg gezeigt, daß mit $\gamma_0 = 1/2(\tilde{\alpha} + \tilde{\beta})$

$$r(\gamma_0 \mathbf{I}, \mathbf{I}) \leq r(\gamma \mathbf{I}, \mathbf{I})$$

für alle $\gamma \in \Gamma_1$ gilt, falls für irgendein $\gamma \in \Gamma_1$

$$r(\gamma, I) < 1$$

ist. Das bedeutet, daß nach Sandberg der Transformationsoperator $\gamma_0 I$ bezüglich aller Transformationsoperatoren $\gamma \in \Gamma_1$ optimal ist. Wir haben in Satz 7 nachgewiesen, daß mit $C_0 = \gamma_0 I$ und $T_0 = I$

$$r(\gamma_0 I, I) \leq r(C, I)$$

für alle $C \in \Gamma_1$ gilt, falls für irgendein $C \in \Gamma_1$

$$r(C, I) < 1$$

ist.

Unter den Voraussetzungen der Gl. (27) sind also unsere optimalen Transformationsoperatoren C_0 und T_0 mit denen von Sandberg identisch. Wir haben aber auf Grund unserer erweiterten Betrachtungsweise mit den Sätzen 5, 7 und 11 die Optimalität von C_0 und T_0 bezüglich einer bedeutend allgemeineren Klasse von Transformationsoperatoren C und T bewiesen.

Der wesentliche Unterschied zwischen unserer Arbeit und der von Sandberg zeigt sich, wenn die Gl. (27) nicht erfüllt ist, d.h. wenn einige der α_i bzw. β_i untereinander verschieden sind. Dieser Fall tritt in der Praxis häufig auf, und wir sind auch für ihn in der Lage, optimale Transformationsoperatoren (im Sinne der Sätze 5, 7 und 11) anzugeben. Grundsätzlich ist auch in diesem Falle Sandbergs Theorie anwendbar und mit unseren Bezeichnungen formulierbar. Dazu müssen wir aber von neuen Vektoren $\hat{\alpha}, \hat{\beta} \in \mathbb{R}^n$ ausgehen, die zwar eine Verschlechterung der diesen Vektoren zugeordneten Kontraktionsschranke $\hat{r}(C, T)$ (bei vorgegebenem C und T) bedeuten (siehe Abschnitt II.2.43), dafür aber die Anwendung der

Sandberg'schen Theorie gemäß Gl. (27) erlauben. Mit

$$\tilde{\alpha} = \min_{i=1, \dots, n} \alpha_i \text{ und } \tilde{\beta} = \max_{i=1, \dots, n} \beta_i \text{ erfüllen die Vektoren}$$

$$\hat{\alpha} = (\tilde{\alpha}, \tilde{\alpha}, \dots, \tilde{\alpha})^T$$

und

$$\hat{\beta} = (\tilde{\beta}, \tilde{\beta}, \dots, \tilde{\beta})^T$$

die geforderten Bedingungen. Mit $\gamma_0 = 1/2(\tilde{\alpha} + \tilde{\beta})$ gilt dann nach Sandberg

$$\gamma_0 \mathbf{l} \in \Gamma_1$$

und

$$\hat{r}(\gamma_0 \mathbf{l}, \mathbf{l}) \leq \hat{r}(\gamma \mathbf{l}, \mathbf{l})$$

für alle $\gamma \mathbf{l} \in \Gamma_1$, falls es irgendein $\gamma \mathbf{l} \in \Gamma_1$ gibt, für das

$$\hat{r}(\gamma \mathbf{l}, \mathbf{l}) < 1$$

ist.

Es stellt sich nun die Frage, ob die in dieser Arbeit vorgeschlagene optimale Kontraktionsschranke $r(\mathbf{C}_0, \mathbf{T}_0)$ grundsätzlich kleiner ist als die optimale Kontraktionsschranke $\hat{r}(\gamma_0 \mathbf{l}, \mathbf{l})$ Sandbergs. Diese Frage muß verneint werden. Nach Satz 11 ist unsere Wahl der Transformationsoperatoren im allgemeinen nicht optimal bezüglich aller $\mathbf{C} \in \Gamma_1$ und $\mathbf{T} \in \Theta$. Wegen

$$r(\gamma_0 \mathbf{l}, \mathbf{l}) = \hat{r}(\gamma_0 \mathbf{l}, \mathbf{l})$$

ist es also durchaus möglich, daß mit der Wahl $\mathbf{C} = \gamma_0 \mathbf{l}$ und $\mathbf{T} = \mathbf{l}$ eine kleinere Kontraktionsschranke erzielt wird und Sandbergs Wahl damit ein besseres Ergebnis liefert. In dem Gegenbeispiel, das für den Beweis von Satz 9 konstruiert wurde, tritt dieser Fall auch tatsächlich ein.

Unter den speziellen Voraussetzungen $T_0 = I$ oder $C_0 = \gamma I$ mit $\gamma \in \mathbf{R}$ kann nachgewiesen werden, daß unsere Kontraktionsschranke sicher nicht größer ist als die Sandbergs:

Satz 20: Es sei $T_0 = I$, und es gebe ein $\gamma I \in \Gamma_1$ derart, daß

$$\hat{r}(\gamma I, I) = r(\gamma I, I) < 1$$

ist. Dann ist $C_0 \in \Gamma_1$, $\gamma_0 I \in \Gamma_1$, und es gilt

$$r(C_0, T_0) \leq \hat{r}(\gamma_0 I, I) < 1.$$

Dabei ist $\hat{r}(\gamma_0 I, I)$ Sandbergs optimale Kontraktions-schranke.

Beweis:

Nach Satz 7 gilt $C_0 \in \Gamma_1$ und

$$r(C_0, T_0) = r(C_0, I) \leq r(\gamma I, I)$$

für alle $\gamma I \in \Gamma_1$. Nach Satz 6 ist $\gamma_0 I \in \Gamma_1$, und wegen

$$\hat{r}(\gamma_0 I, I) = r(\gamma_0 I, I)$$

folgt schließlich

$$r(C_0, T_0) \leq r(\gamma_0 I, I) = \hat{r}(\gamma_0 I, I).$$

q.e.d.

Satz 21: Es sei $C_0 = \gamma I \in \Gamma_1$ mit $\gamma = 1/2(\alpha_i + \beta_i)$ für $i=1, \dots, n$, dann gilt

$$r(C_0, T_0) \leq \hat{r}(\gamma_0 I, I),$$

falls $\gamma_0 I \in \Gamma_1$ ist. Dabei ist $\hat{r}(\gamma_0 I, I)$ Sandbergs optimale Kontraktionsschranke.

Beweis:

Wegen Satz 5 gilt

$$r(\mathbf{C}_0, \mathbf{T}_0) \leq r(\mathbf{C}_0, \mathbf{I}) = r(\mathbf{YI}, \mathbf{I}).$$

Ferner ist $\mathbf{Y} = \mathbf{Y}_0$, und mit

$$r(\mathbf{Y}_0 \mathbf{I}, \mathbf{I}) = \hat{r}(\mathbf{Y}_0 \mathbf{I}, \mathbf{I})$$

folgt schließlich

$$r(\mathbf{C}_0, \mathbf{T}_0) \leq r(\mathbf{Y}_0 \mathbf{I}, \mathbf{I}) = \hat{r}(\mathbf{Y}_0 \mathbf{I}, \mathbf{I}).$$

q.e.d.

Satz 22: Es sei $\mathbf{C}_0 = \mathbf{YI}$ mit $\mathbf{Y} = 1/2(\mathbf{a}_i + \mathbf{\beta}_i)$ für $i=1, \dots, n$.
Ferner gelte $\mathbf{Y}_0 \mathbf{I} \in \mathbf{F}_1$ und

$$\hat{r}(\mathbf{Y}_0 \mathbf{I}, \mathbf{I}) < 1.$$

Dann ist $\mathbf{C}_0 \in \mathbf{F}_1$ und

$$r(\mathbf{C}_0, \mathbf{T}_0) \leq \hat{r}(\mathbf{Y}_0 \mathbf{I}, \mathbf{I}).$$

Dabei ist $\hat{r}(\mathbf{Y}_0 \mathbf{I}, \mathbf{I})$ Sandbergs optimale Kontraktions-
schranke.

Beweis:

Wegen

$$\hat{r}(\mathbf{Y}_0 \mathbf{I}, \mathbf{I}) = r(\mathbf{Y}_0 \mathbf{I}, \mathbf{I}) < 1$$

gilt nach Satz 6 $\mathbf{C}_0 \in \mathbf{F}_1$. Die Behauptung

$$r(\mathbf{C}_0, \mathbf{T}_0) \leq \hat{r}(\mathbf{Y}_0 \mathbf{I}, \mathbf{I})$$

folgt aus Satz 7.

q.e.d.

Grundsätzlich kann gesagt werden, daß Sandbergs Wahl der Transformationsoperatoren nur unter seinen speziellen Voraussetzungen zu rechtfertigen ist. Unter unseren allgemeineren Voraussetzungen bezüglich des nichtlinearen Operators $\mathbf{N}$ ist Sandbergs Wahl nur dann zu begründen, wenn man einen Teil der Information über $\mathbf{N}$ aufgibt. Unter diesen weniger informativen Voraussetzungen kommen wir mit unserer Theorie zum selben Ergebnis wie Sandberg. Durch eine subtilere Betrachtungsweise sind wir aber zu diesem Informationsverlust nicht gezwungen und erhalten mit Satz 7 eine Begründung für die Wahl $\mathbf{C} = \mathbf{C}_0$ und $\mathbf{T} = \mathbf{T}_0$, die als direkte Verallgemeinerung von Sandbergs Ergebnis (bezüglich seiner speziellen Voraussetzungen) verstanden werden kann. Diese Wahl wird in den Sätzen 5 und 11 weiter erhärtet.

Im Lichte dieser Sätze erscheint jede von $\mathbf{C}_0$ und $\mathbf{T}_0$ verschiedene a priori-Wahl der Transformationsoperatoren $\mathbf{C}$ und $\mathbf{T}$ als willkürlich.

2.4.4 Stetigkeitsbetrachtungen

Wir nehmen an, daß alle Voraussetzungen und die Bedingungen (i), (ii) und (iii) von Satz 2 erfüllt sind. Dann existieren auf Grund dieses Satzes für alle Erregungen $\mathbf{e} \in \mathbf{B}^n$ eindeutige Systemreaktionen $\mathbf{f} \in \mathbf{B}^n$, welche die Operatorgleichungen

$$\mathbf{e} = \mathbf{f} + \mathbf{H}\mathbf{N}\mathbf{f}$$

befriedigen. Anders ausgedrückt: Der nichtlineare Systemoperator

$$(\mathbf{I} + \mathbf{H}\mathbf{N})^{-1}: \mathbf{B}^n \longrightarrow \mathbf{B}^n$$

existiert, und es ist

$$\mathbf{f} = (\mathbf{I} + \mathbf{H}\mathbf{N})^{-1}\mathbf{e}.$$

Es sei mit $e_1, e_2 \in \mathbb{R}^n$

$$e_1 = f_1 + \mathbf{H}\mathbf{N}f_1$$

und

$$e_2 = f_2 + \mathbf{H}\mathbf{N}f_2.$$

Nach der gemäß Satz 2 durchgeführten Transformation gilt

$$(\mathbf{I} + \mathbf{H}\mathbf{C})^{-1}e_1 = f_1 + (\mathbf{I} + \mathbf{H}\mathbf{C})^{-1}\mathbf{H}(\mathbf{N} - \mathbf{C})f_1$$

und

$$(\mathbf{I} + \mathbf{H}\mathbf{C})^{-1}e_2 = f_2 + (\mathbf{I} + \mathbf{H}\mathbf{C})^{-1}\mathbf{H}(\mathbf{N} - \mathbf{C})f_2.$$

Nach Differenzbildung beider Gleichungen folgt

$$(\mathbf{I} + \mathbf{H}\mathbf{C})^{-1}(e_1 - e_2) = (f_1 - f_2) + (\mathbf{I} + \mathbf{H}\mathbf{C})^{-1}\mathbf{H}\mathbf{T}^{-1}(\mathbf{T}(\mathbf{N} - \mathbf{C})f_1 - \mathbf{T}(\mathbf{N} - \mathbf{C})f_2)$$

und damit

$$\|(\mathbf{I} + \mathbf{H}\mathbf{C})^{-1}(e_1 - e_2)\| \geq \|f_1 - f_2\| - r(\mathbf{C}, \mathbf{T}) \cdot \|f_1 - f_2\|$$

bzw.

$$\|f_1 - f_2\| \leq \frac{\|(\mathbf{I} + \mathbf{H}\mathbf{C})^{-1}\|}{1 - r(\mathbf{C}, \mathbf{T})} \|e_1 - e_2\|.$$

Das System ist also lipschitzstetig, und für die inkrementale Verstärkung des Systemoperators ergibt sich

$$g_0((\mathbf{I} + \mathbf{H}\mathbf{N})^{-1}) \leq \frac{\|(\mathbf{I} + \mathbf{H}\mathbf{C})^{-1}\|}{1 - r(\mathbf{C}, \mathbf{T})}.$$

Analoge Überlegungen führen bei Erfülltsein der Voraussetzungen und Bedingungen von Satz 3 auf

$$g_0((\mathbf{N} + \mathbf{H})^{-1}) \leq \frac{\|(\mathbf{C} + \mathbf{H})^{-1}\|}{1 - q(\mathbf{C}, \mathbf{T})},$$

womit die Lipschitzstetigkeit des Systemoperators $(\mathbf{N} + \mathbf{H})^{-1}$ nachgewiesen ist.

Grundsätzlich kann gesagt werden, daß unter den Bedingungen für die eindeutige Lösbarkeit der Operatorgleichungen die diesen Gleichungen zugeordneten Systemoperatoren Lipschitzstetig sind. Anschaulich bedeutet dies, daß kleine Erregungsänderungen immer kleine Änderungen in den Systemreaktionen bewirken.

2.4.5 Allgemeines

Aus den Sätzen 2 bzw. 3 ist ersichtlich, daß mit dem Erfülltsein ihrer Voraussetzungen und Bedingungen nicht nur die Existenz einer Lösung garantiert wird, sondern auch ein Verfahren zu deren Berechnung angegeben werden kann. Um dieses Verfahren anwenden zu können, werden zunächst die Gl. (1) bzw. (2) in folgende, jeweils äquivalente Operatorgleichungen transformiert. (Auf Grund der Bedingungen in den Sätzen existieren die Inversen)

$$f = -(I + HC)^{-1}H(N - C)f + (I + HC)^{-1}e \quad (28)$$

$$f = -(C + H)^{-1}(N - C)f + (C + H)^{-1}e. \quad (29)$$

Diese Gleichungen entstehen dadurch, daß man vom nichtlinearen Operator **N** einen gewissen linearen Operator **C** abzieht und diesen gleichzeitig dem linearen Operator **H** wieder zufügt. Wäre nun **N = C**, so hätte man bereits die Lösungen

$$f_L = (I + HC)^{-1}e$$

bzw.

$$f_L = (C + H)^{-1}e,$$

die als Lösungen der durch $\mathbf{N}=\mathbf{C}$ linearisierten Gleichungen verstanden werden können. Die Invertierbarkeit der Operatoren $(\mathbf{I}+\mathbf{H}\mathbf{C})$ bzw. $(\mathbf{C}+\mathbf{H})$ bedeutet also, daß die durch $\mathbf{N}=\mathbf{C}$ linearisierten Ersatzsysteme eindeutige Lösungen besitzen. Unter Verwendung dieser Ersatzsysteme können die Blockschaltbilder der Abb. (1) bzw. (2) durch die äquivalenten Blockschaltbilder der Abb. (9) bzw. (10) gemäß den Gleichungen (28) bzw. (29) ersetzt werden.

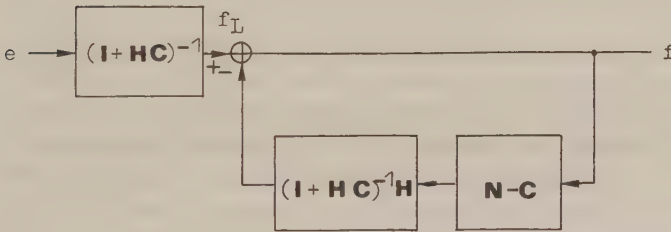


Abb. (9)

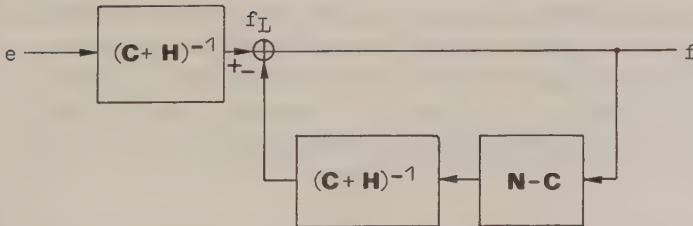


Abb. (10)

Die grundsätzliche Form der transformierten Gleichungen ist

$$f = -\tilde{H}\tilde{N}f + f_L,$$

und da nach den Bedingungen (iii) der Sätze 2 bzw. 3 der Operator $\tilde{H}\tilde{N}$ eine Kontraktion ist, existieren auch eindeutige Lösungen der nichtlinearen Operatorgleichungen, und sie können anhand der Iterationsvorschrift

$$f_{\mu+1} = -\tilde{H}\tilde{N}f_{\mu} + f_L \quad \mu=0,1,2,\dots$$

unter Verwendung eines beliebigen Startwertes $f_0 \in B^n$ mit beliebiger Genauigkeit berechnet werden.

Will man eine Gleichung der Bauart Gl. (1) bzw. Gl. (2) auf ihre Lösbarkeit hin untersuchen, so stellt sich die Frage, nach welchen Gesichtspunkten man die Transformationsoperatoren $\mathbf{C}$ und $\mathbf{T}$ wählen soll, anhand derer man dann die Bedingungen in den Sätzen 2 bzw. 3 überprüft. Da eine wiederholte Überprüfung mit immer neu gewählten Operatoren $\mathbf{C}$ und $\mathbf{T}$ willkürlich und zu aufwendig wäre, wird man versuchen, die Bedingungen mit gezielt gewählten Operatoren zu testen. Wie im Abschnitt II.2.4.3 gezeigt wurde, ist es sinnvoll, den Test mit der Wahl $\mathbf{C}=\mathbf{C}_0$ und $\mathbf{T}=\mathbf{T}_0$ durchzuführen. Falls nämlich die Bedingungen der Sätze 2 bzw. 3 mit diesen optimalen Transformationsoperatoren nicht erfüllt sind, sind sie es auch nicht bezüglich einer sehr allgemeinen Klasse von Operatoren. Aus diesem Grunde wollen wir im folgenden immer diese optimalen Transformationsoperatoren $\mathbf{C}_0$ und $\mathbf{T}_0$ verwenden.

2.5 Spezielle Bedingungen für die Lösbarkeit der Operatorgleichungen

Nachdem wir das Problem der Existenz von Lösungen der Operatorgleichungen (1) und (2) unter den sehr allgemeinen Voraussetzungen der Abschnitte II.2.1, II.2.2 und II.2.4 behandelt haben, wollen wir uns jetzt mit einschränkenden Voraussetzungen begnügen. Es wird sich zeigen, daß man interessante und gut interpretierbare hinreichende Existenzbedingungen erhält, wenn man als Banachraum $\mathbf{B}^n$ den Hilbertraum $\mathbf{L}_{2r}^n(-\infty, \infty)$ wählt. Weiter wollen wir annehmen, daß der nichtlineare Operator $\mathbf{N}$ ein Element der Klasse $\Psi(\alpha, \beta)$ und $\mathbf{H}$ ein linearer Faltungsoperator gemäß Abschnitt I.4.4 ist.

Es wird im folgenden unsere Aufgabe sein, die Voraussetzungen der Sätze 2 und 3 im Hinblick auf die hier getroffenen Vereinbarungen neu zu formulieren.

2.5.1 Voraussetzungen und Vereinbarungen

Als Banachraum $\mathbf{B}^n$ wählen wir den reellen Hilbertraum $\mathbf{L}_{2r}^n(-\infty, \infty)$ (siehe Bemerkung 3 in Abschnitt II.2.1). Wir setzen weiter voraus, daß $\mathbf{N} \in \Psi(\alpha, \beta)$ ist. Ferner sei $\mathbf{H}$ ein linearer Faltungsoperator (siehe Abschnitt I.4.4), derart, daß die Gleichung $f = \mathbf{H}g$ mit $g \in \mathbf{L}_{2r}^n(-\infty, \infty)$ durch

$$f(t) = h_0 g(t) + \int_{-\infty}^{\infty} h(t-\tau)g(\tau)d\tau \text{ für alle } t \in \mathbf{R}$$

mit der reellen $n \times n$ -Matrix h_0 und $h \in \mathbf{M}_{1r}^{n,n}(-\infty, \infty)$ definiert ist.

Damit ist auch $f \in L_{2r}^n(-\infty, \infty)$. Der Operator H kann im Frequenzbereich durch

$$H(j\omega) = h_0 + \int_{-\infty}^{\infty} h(t) e^{-j\omega t} dt \quad \text{für alle } \omega \in \mathbb{R}$$

und

$$F(j\omega) = H(j\omega)G(j\omega) \quad \text{für alle } \omega \in \mathbb{R}$$

beschrieben werden, wobei $F(j\omega)$ bzw. $G(j\omega)$ die L_2 -Fourier-Transformierten von f bzw. g sind (siehe Abschnitt I.4.4). Mit der Abkürzung

$$\Lambda\{M\} = \sqrt{\lambda_{\max}(M^H M)}$$

gilt nach Abschnitt I.4.4 wie in Abschnitt II.2.2 gefordert

$$\|H\| = \sup_{\omega \in \mathbb{R}} \Lambda\{H(j\omega)\}.$$

Den optimalen Transformationsoperatoren C_0 und T_0 des Abschnittes II.2.4.3 werden die durch

$$C_0 = \text{diag}(c_1^{(0)}, c_2^{(0)}, \dots, c_n^{(0)})$$

und

$$T_0 = \text{diag}(t_1^{(0)}, t_2^{(0)}, \dots, t_n^{(0)})$$

definierten Matrizen C_0 und T_0 zugeordnet, und wir setzen

$$\eta(C_0, T_0) = \eta(C_0, T_0).$$

Bemerkung 1:

Nach Konstruktion des Operators T_0 existiert die Inverse der Matrix T_0 .

Bemerkung 2:

Die Anwendung des Operators $\mathbf{C}_0$ auf irgendein $g \in L_{2r}^n(-\infty, \infty)$ kann mit $f = \mathbf{C}_0 g$ im Zeitbereich durch

$$f(t) = (\mathbf{C}_0 g)(t) = C_0 g(t) \text{ für alle } t \in \mathbf{R}$$

bzw. im Frequenzbereich durch

$$F(j\omega) = C_0 G(j\omega) \quad \text{für alle } \omega \in \mathbf{R}$$

definiert werden. (Das Analoge gilt für $\mathbf{T}_0$.)

2.5.2 Bedingungen für die Existenz der Operatoren $(\mathbf{I} + \mathbf{H}\mathbf{C}_0)^{-1}$ und $(\mathbf{C}_0 + \mathbf{H})^{-1}$

Die Frage nach den Bedingungen für die Existenz von $(\mathbf{I} + \mathbf{H}\mathbf{C}_0)^{-1}$ und $(\mathbf{C}_0 + \mathbf{H})^{-1}$ soll unter den Voraussetzungen des Abschnitts II.2.5.1 durch folgende Sätze beantwortet werden:

Satz 23: Der lineare Operator

$$(\mathbf{I} + \mathbf{H}\mathbf{C}_0) : L_{2r}^n(-\infty, \infty) \longrightarrow L_{2r}^n(-\infty, \infty)$$

besitzt eine beschränkte Inverse, falls die Bedingungen

(i) $(\mathbf{I}_n + \mathbf{H}(j\omega)\mathbf{C}_0)^{-1}$ existiert für alle $\omega \in \mathbf{R}$ und

(ii) $\sup_{\omega \in \mathbf{R}} \Lambda \{(\mathbf{I}_n + \mathbf{H}(j\omega)\mathbf{C}_0)^{-1}\} < \infty$

erfüllt sind. Dann gilt ferner

$$\|(\mathbf{I} + \mathbf{H}\mathbf{C}_0)^{-1}\| = \sup_{\omega \in \mathbf{R}} \Lambda \{(\mathbf{I}_n + \mathbf{H}(j\omega)\mathbf{C}_0)^{-1}\} \quad (30)$$

und

$$\begin{aligned} \|(\mathbf{I} + \mathbf{H}\mathbf{C}_0)^{-1} \mathbf{H} \mathbf{T}_0^{-1}\| &= \\ &= \sup_{\omega \in \mathbf{R}} \Lambda \{(\mathbf{I}_n + \mathbf{H}(j\omega)\mathbf{C}_0)^{-1} \mathbf{H}(j\omega)\mathbf{T}_0^{-1}\}. \end{aligned} \quad (31)$$

Beweis:

Zur Abkürzung schreiben wir $R(j\omega) = 1_{n+H(j\omega)} C_0$. Die Elemente der Matrix $R(j\omega)$ sind nach Abschnitt I.4.4 in ω gleichmäßig stetig. Wegen der Voraussetzung (i) des Satzes ist

$$\det R(j\omega) \neq 0 \text{ für alle } \omega \in \mathbf{R}.$$

Alle Elemente der Matrix

$$R^{-1}(j\omega) = \frac{1}{\det R(j\omega)} \operatorname{adj} R(j\omega)$$

sind stetige Funktionen von ω , da sowohl $\det R(j\omega)$ als auch alle Elemente von $\operatorname{adj} R(j\omega)$ in ω stetig sind und ferner die Determinante nicht verschwindet. Daraus folgt mit Voraussetzung (ii)

$$\operatorname{ess\,sup}_{\omega \in \mathbf{R}} \Lambda \{R^{-1}(j\omega)\} = \sup_{\omega \in \mathbf{R}} \Lambda \{R^{-1}(j\omega)\} < \infty. \quad (32)$$

Es sei $G(j\omega)$ die $\mathbf{L}_2$ -Fourier-Transformierte eines beliebigen $g \in \mathbf{L}_{2\mathbf{R}}^n(-\infty, \infty)$. Dann ist die Funktion

$$\hat{F}(\omega) = F(j\omega) = R^{-1}(j\omega)G(j\omega) \quad \text{für alle } \omega \in \mathbf{R} \quad (33)$$

meßbar, und es gilt

$$\begin{aligned} & \int_{-\infty}^{\infty} G^H(j\omega) R^{-1H}(j\omega) R^{-1}(j\omega) G(j\omega) d\omega \\ & \leq \int_{-\infty}^{\infty} \Lambda^2 \{R^{-1}(j\omega)\} G^H(j\omega) G(j\omega) d\omega \\ & \leq \operatorname{ess\,sup}_{\omega \in \mathbf{R}} \Lambda^2 \{R^{-1}(j\omega)\} \cdot \int_{-\infty}^{\infty} G^H(j\omega) G(j\omega) d\omega \\ & = \sup_{\omega \in \mathbf{R}} \Lambda^2 \{R^{-1}(j\omega)\} \int_{-\infty}^{\infty} G^H(j\omega) G(j\omega) d\omega < \infty. \end{aligned}$$

Die Funktion F ist also in $L_2^n(-\infty, \infty)$ enthalten. Somit existiert nach Abschnitt I.4.2 ein eindeutiges $f \in L_2^n(-\infty, \infty)$ mit der L_2 -Fourier-Transformierten $F(j\omega)$. Insbesondere ist $f \in L_{2r}^n(-\infty, \infty)$, da

$$R(j\omega) = R^*(-j\omega) \quad \text{für alle } \omega \in \mathbf{R}$$

und damit

$$R^{-1}(j\omega) = R^{-1*}(-j\omega) \quad \text{für alle } \omega \in \mathbf{R} \quad (34)$$

gilt.

Aus dem Vorhergehenden folgt, daß für jedes $g \in L_{2r}^n(-\infty, \infty)$ ein eindeutiges $f \in L_{2r}^n(-\infty, \infty)$ existiert, derart, daß

$$G(j\omega) = (I_n + H(j\omega)C_0)F(j\omega) \quad \text{für alle } \omega \in \mathbf{R}$$

bzw. $g = (I + HC_0)f$ gilt.

Damit existiert der Operator $(I + HC_0)^{-1}$. Wegen Gl. (32), (33) und (34) sind die Voraussetzungen von Satz I.14 erfüllt. Folglich ist $(I + HC_0)^{-1}$ beschränkt, und für die Norm gilt Gl. (30). Ähnliche Überlegungen führen auf Gl. (31).

q.e.d.

Satz 24: Der lineare Operator

$$(C_0 + H) : L_{2r}^n(-\infty, \infty) \longrightarrow L_{2r}^n(-\infty, \infty)$$

besitzt eine beschränkte Inverse, falls die Bedingungen

(i) $(C_0 + H(j\omega))^{-1}$ existiert für alle $\omega \in \mathbf{R}$ und

(ii) $\sup_{\omega \in \mathbf{R}} \Lambda \{ (C_0 + H(j\omega))^{-1} \} < \infty$

erfüllt sind.

Dann gilt ferner

$$\|(\mathbf{C}_0 + \mathbf{H})^{-1}\| = \sup_{\omega \in \mathbf{R}} \wedge \{(\mathbf{C}_0 + \mathbf{H}(j\omega))^{-1}\}$$

und

$$\|(\mathbf{C}_0 + \mathbf{H})^{-1} \mathbf{T}^{-1}\| = \sup_{\omega \in \mathbf{R}} \wedge \{(\mathbf{C}_0 + \mathbf{H}(j\omega))^{-1} \mathbf{T}_0^{-1}\}.$$

Beweis:

Analog zum Beweis von Satz 23.

In den folgenden Sätzen werden die Voraussetzungen (i) und (ii) der Sätze 23 bzw. 24 anders formuliert.

Satz 25: Der lineare Operator

$$(\mathbf{I} + \mathbf{H}\mathbf{C}_0): \mathbf{L}_{2\mathbf{r}}^n(-\infty, \infty) \longrightarrow \mathbf{L}_{2\mathbf{r}}^n(-\infty, \infty)$$

besitzt eine beschränkte Inverse, falls die Bedingungen

(i) $\det(\mathbf{I}_{n+H(j\omega)\mathbf{C}_0}) \neq 0$ für alle $\omega \in \mathbf{R}$ und

(ii) $\det(\mathbf{I}_{n+H_0\mathbf{C}_0}) \neq 0$

erfüllt sind. Dann gilt ferner

$$\|(\mathbf{I} + \mathbf{H}\mathbf{C}_0)^{-1}\| = \sup_{\omega \in \mathbf{R}} \wedge \{(\mathbf{I}_{n+H(j\omega)\mathbf{C}_0})^{-1}\}$$

und

$$\|(\mathbf{I} + \mathbf{H}\mathbf{C}_0)^{-1} \mathbf{H} \mathbf{T}_0^{-1}\| = \sup_{\omega \in \mathbf{R}} \wedge \{(\mathbf{I}_{n+H(j\omega)\mathbf{C}_0})^{-1} \mathbf{H}(j\omega) \mathbf{T}_0^{-1}\}.$$

Beweis:

Wir wollen zeigen, daß aus den Bedingungen (i) und (ii) dieses Satzes die Bedingungen (i) und (ii) von Satz 23 folgen. Die Bedingungen (i) beider Sätze sind offenbar äquivalent.

Es bleibt also zu zeigen, daß

$$\sup_{\omega \in \mathbb{R}} \Lambda \{ (1_{\mathbb{N}} + H(j\omega)C_0)^{-1} \} < \infty$$

ist. Wegen $h \in M_{1 \times \mathbb{R}}^{\mathbb{N}, \mathbb{N}}(-\infty, \infty)$ gilt

$$\lim_{|\omega| \rightarrow \infty} H(j\omega) = h_0.$$

Damit strebt $\det (1_{\mathbb{N}} + H(j\omega)C_0)$ für $|\omega| \rightarrow \infty$ gegen $\det (1_{\mathbb{N}} + h_0 C_0)$.
Ferner ist $\det (1_{\mathbb{N}} + H(j\omega)C_0)$ eine in ω gleichmäßig stetige Funktion. Folglich implizieren die Voraussetzungen (i) und (ii), daß

$$\inf_{\omega \in \mathbb{R}} |\det (1_{\mathbb{N}} + H(j\omega)C_0)| > 0$$

ist, und da alle Elemente von $1_{\mathbb{N}} + H(j\omega)C_0$ gleichmäßig absolut beschränkt sind, gilt schließlich

$$\sup_{\omega \in \mathbb{R}} \Lambda \{ (1_{\mathbb{N}} + H(j\omega)C_0)^{-1} \} < \infty.$$

Es läßt sich einfach zeigen, daß auch aus den Voraussetzungen von Satz 23 die Voraussetzungen dieses Satzes folgen. Das bedeutet also, daß die Voraussetzungen beider Sätze äquivalent sind.

q.e.d.

Satz 26: Der lineare Operator

$$(C_0 + H): L_{2\mathbb{R}}^{\mathbb{N}}(-\infty, \infty) \longrightarrow L_{2\mathbb{R}}^{\mathbb{N}}(-\infty, \infty)$$

besitzt eine beschränkte Inverse, falls die Bedingungen

(i) $\det (C_0 + H(j\omega)) \neq 0$ für alle $\omega \in \mathbb{R}$ und

(ii) $\det (C_0 + h_0) \neq 0$

erfüllt sind.

Dann gilt ferner

$$\|(\mathbf{C}_0 + \mathbf{H})^{-1}\| = \sup_{\omega \in \mathbf{R}} \Lambda \{(\mathbf{C}_0 + \mathbf{H}(j\omega))^{-1}\}$$

und

$$\|(\mathbf{C}_0 + \mathbf{H})^{-1} \mathbf{T}_0^{-1}\| = \sup_{\omega \in \mathbf{R}} \Lambda \{(\mathbf{C}_0 + \mathbf{H}(j\omega))^{-1} \mathbf{T}_0^{-1}\}.$$

Beweis:

Der Beweis erfolgt analog zum Beweis von Satz 25.

2.5.3 Lösung der Gleichungen $e = f + \mathbf{H}\mathbf{N}f$ und $e = \mathbf{N}f + \mathbf{H}f$

Mit den speziellen Voraussetzungen des Abschnitts II.2.5 sind wir jetzt in der Lage, die Sätze 2 und 3 neu zu formulieren. Damit erhalten wir die folgenden Sätze, welche z.B. in der Netzwerktheorie von Bedeutung sind. Wir wollen dabei alle Voraussetzungen explizit angeben, damit die Sätze in sich abgeschlossen sind.

Satz 27: In der nichtlinearen Operatorgleichung $e = f + \mathbf{H}\mathbf{N}f$ sei der lineare Operator $\mathbf{H}$ ein Faltungsoperator und der nichtlineare Operator $\mathbf{N}$ ein Element von $\Psi(\alpha, \beta)$. Gemäß den Abschnitten I.4.4 und II.2.3 gilt also für alle $t \in \mathbf{R}$

$$e(t) = f(t) + h_0 \psi(f(t), t) + \int_{-\infty}^{\infty} h(t-\tau) \psi(f(\tau), \tau) d\tau,$$

wobei h_0 eine reelle $n \times n$ -Matrix und $h \in \mathbf{M}_{1r}^{n,n}(-\infty, \infty)$ ist. Ferner sei

$$H(j\omega) = h_0 + \int_{-\infty}^{\infty} h(t) e^{-j\omega t} dt \text{ für alle } \omega \in \mathbf{R}.$$

Falls die Bedingungen

- (i) $(1_{\mathbf{N}} + \mathbf{H}(\mathbf{j}\omega)\mathbf{C}_0)^{-1}$ existiert für alle $\omega \in \mathbf{R}$,
 - (ii) $\sup_{\omega \in \mathbf{R}} \wedge \{ (1_{\mathbf{N}} + \mathbf{H}(\mathbf{j}\omega)\mathbf{C}_0)^{-1} \} < \infty$ und
 - (iii) $r_0 = \sup_{\omega \in \mathbf{R}} \wedge \{ (1_{\mathbf{N}} + \mathbf{H}(\mathbf{j}\omega)\mathbf{C}_0)^{-1} \mathbf{H}(\mathbf{j}\omega) \mathbf{T}_0^{-1} \} \cdot \eta(\mathbf{C}_0, \mathbf{T}_0) < 1$
- erfüllt sind, gilt:

Für jedes $e \in \mathbf{L}_{2\mathbf{r}}^{\mathbf{n}}(-\infty, \infty)$ existiert ein eindeutiges $f \in \mathbf{L}_{2\mathbf{r}}^{\mathbf{n}}(-\infty, \infty)$ derart, daß $e = f + \mathbf{H}\mathbf{N}f$ gilt. Die Lösungsfunktion f kann anhand der Iterationsvorschrift $f = \lim_{\mu \rightarrow \infty} f_{\mu}$ mit

$$f_{\mu+1} = -(\mathbf{I} + \mathbf{H}\mathbf{C}_0)^{-1} \mathbf{H}(\mathbf{N} - \mathbf{C}_0) f_{\mu} + (\mathbf{I} + \mathbf{H}\mathbf{C}_0)^{-1} e$$

ermittelt werden, wobei f_0 ein beliebiges Element aus $\mathbf{L}_{2\mathbf{r}}^{\mathbf{n}}(-\infty, \infty)$ ist. Für die μ -te Approximation gilt

$$\|f - f_{\mu}\| \leq \frac{r_0^{\mu}}{1 - r_0} \|f_1 - f_0\|.$$

Beweis:

Es ist zu beweisen, daß unter den genannten Bedingungen die Voraussetzungen des Satzes 2 für $\mathbf{B}^{\mathbf{n}} = \mathbf{L}_{2\mathbf{r}}^{\mathbf{n}}(-\infty, \infty)$ erfüllt sind. Nach der Bemerkung von Abschnitt II.2.4.3 existiert $\mathbf{T}_0^{-1}$. Aus (i) und (ii) folgt mit Satz 23 die Invertierbarkeit von $(\mathbf{I} + \mathbf{H}\mathbf{C}_0)$. Ferner folgt aus Satz 23, daß

$$\begin{aligned} r(\mathbf{C}_0, \mathbf{T}_0) &= \|(\mathbf{I} + \mathbf{H}\mathbf{C}_0)^{-1} \mathbf{H} \mathbf{T}_0^{-1}\| \cdot \eta(\mathbf{C}_0, \mathbf{T}_0) \\ &= \sup_{\omega \in \mathbf{R}} \wedge \{ (1_{\mathbf{N}} + \mathbf{H}(\mathbf{j}\omega)\mathbf{C}_0)^{-1} \mathbf{H}(\mathbf{j}\omega) \mathbf{T}_0^{-1} \} \cdot \eta(\mathbf{C}_0, \mathbf{T}_0) \end{aligned}$$

gilt, und mit Bedingung (iii) folgt schließlich

$$r(\mathbf{C}_0, \mathbf{T}_0) = r_0 < 1.$$

q.e.d.

Satz 28: In der nichtlinearen Operatorgleichung $e = \mathbf{N}f + \mathbf{H}f$ sei der lineare Operator $\mathbf{H}$ ein Faltungsoperator und der nichtlineare Operator $\mathbf{N}$ ein Element von $\Psi(\alpha, \beta)$. Gemäß den Abschnitten I.4.4 und II.2.3 gilt also für alle $t \in \mathbf{R}$

$$e(t) = \psi(f(t), t) + \int_{-\infty}^{\infty} h(t-\tau) f(\tau) d\tau + h_0 f(t),$$

wobei h_0 eine reelle $n \times n$ -Matrix und $h \in \mathbf{M}_{1r}^{n,n}(-\infty, \infty)$ ist. Ferner sei

$$H(j\omega) = h_0 + \int_{-\infty}^{\infty} h(t) e^{j\omega t} dt \text{ für alle } \omega \in \mathbf{R}.$$

Falls die Bedingungen

- (i) $(C_0 + H(j\omega))^{-1}$ existiert für alle $\omega \in \mathbf{R}$,
- (ii) $\sup_{\omega \in \mathbf{R}} \Lambda \{ (C_0 + H(j\omega))^{-1} \} < \infty$ und
- (iii) $q_0 = \sup_{\omega \in \mathbf{R}} \Lambda \{ (C_0 + H(j\omega))^{-1} T_0^{-1} \} \cdot \eta(C_0, T_0) < 1$

erfüllt sind, gilt:

Für jedes $e \in \mathbf{L}_{2r}^n(-\infty, \infty)$ existiert ein eindeutiges $f \in \mathbf{L}_{2r}^n(-\infty, \infty)$ derart, daß $e = \mathbf{N}f + \mathbf{H}f$ gilt. Die Lösungsfunktion f kann anhand der Iterationsvorschrift $f = \lim_{\mu \rightarrow \infty} f_\mu$ mit

$$f_{\mu+1} = -(C_0 + H)^{-1} (N - C_0) f_\mu + (C_0 + H)^{-1} e$$

ermittelt werden, wobei f_0 ein beliebiges Element aus $\mathbf{L}_{2r}^n(-\infty, \infty)$ ist. Für die μ -te Approximation gilt

$$\|f - f_\mu\| \leq \frac{q_0^\mu}{1 - q_0} \|f_1 - f_0\|.$$

Beweis:

Der Beweis verläuft analog zu dem von Satz 27.

Bemerkung:

Die Voraussetzungen (i) und (ii) der Sätze 27 bzw. 28 können durch die äquivalenten Bedingungen (i) und (ii) der Sätze 25 bzw. 26 ersetzt werden.

2.5.3.1 Die Operatorgleichungen bei passivem linearen und nichtlinearen Systemanteil

Unter den Voraussetzungen von Abschnitt II.2.5.1 ist der lineare Operator $\mathbf{H}$ passiv, falls

$$\mathbf{H}^H(j\omega) + \mathbf{H}(j\omega) \geq 0 \quad \text{für alle } \omega \in \mathbb{R}$$

gilt. Wir nennen den nichtlinearen Operator $\mathbf{N} \in \Psi(\alpha, \beta)$ passiv, falls $\alpha_i > 0$ für $i = 1, \dots, n$ ist. Das bedeutet, daß die nichtlinearen Funktionen $\psi_i(u, t)$ bezüglich u zu jedem "Zeitpunkt" t streng monoton steigend sind (siehe z.B. Abb. (11)).

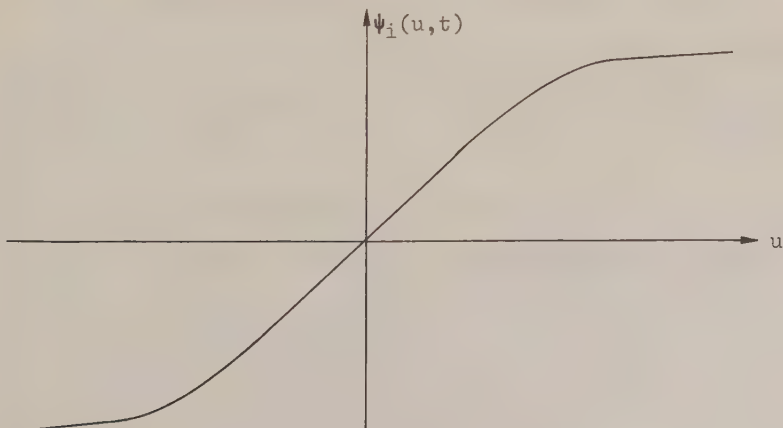


Abb. (11)

Für Systeme dieser Art gilt der folgende bemerkenswerte Satz:

Satz 29: Es seien die Voraussetzungen von Satz 27 gegeben.

Falls die Bedingungen

$$(i) \quad a_i > 0 \text{ für } i = 1, \dots, n,$$

$$(ii) \quad H^H(j\omega) + H(j\omega) \geq 0 \text{ für alle } \omega \in \mathbf{R} \text{ und}$$

$$(iii) \quad \mathbf{T}_0 = \mathbf{I} \text{ oder } \mathbf{C}_0 = \gamma \mathbf{I} \text{ mit } \gamma \in \mathbf{R}$$

erfüllt sind, gelten auch die Bedingungen (i), (ii) und (iii) von Satz 27.

Beweis:

Die Bedingungen (i) und (ii) von Satz 27 sind nach Satz 25 und dessen Beweis erfüllt, falls

$$\det (\mathbf{1}_n + H(j\omega) \mathbf{C}_0) \neq 0 \text{ für alle } \omega \in \mathbf{R} \text{ und}$$

$$\det (\mathbf{1}_n + h_0 \mathbf{C}_0) \neq 0$$

gilt. Aus $c_i^{(0)} = \frac{1}{2}(a_i + \beta_i)$ für $i = 1, \dots, n$ folgt $C_0 > 0$, C_0^{-1} existiert und $C_0^{-1} > 0$. Wegen

$$\det (\mathbf{1}_n + H(j\omega) \mathbf{C}_0) = \det (C_0^{-1} + H(j\omega)) \cdot \det C_0$$

und

$$\det (\mathbf{1}_n + h_0 \mathbf{C}_0) = \det (C_0^{-1} + h_0) \cdot \det C_0$$

genügt es zu zeigen, daß

$$\det (C_0^{-1} + H(j\omega)) \neq 0 \text{ für alle } \omega \in \mathbf{R}$$

und

$$\det (C_0^{-1} + h_0) \neq 0$$

gilt.

Wir nehmen an, daß für irgendein $\omega_0 \in \mathbf{R}$

$$\det (C_0^{-1} + H(j\omega_0)) = 0$$

ist. Dann gibt es ein $x \in \mathbf{C}^n$ mit $x \neq 0$, für das

$$x^H (C_0^{-1} + H(j\omega_0)) x = 0$$

gilt. Folglich ist

$$\begin{aligned} x^H (2C_0^{-1} + H^H(j\omega_0) + H(j\omega_0)) x &= \\ 2x^H C_0^{-1} x + x^H (H^H(j\omega_0) + H(j\omega_0)) x &= 0, \end{aligned}$$

was aber nicht sein kann, da $C_0^{-1} > 0$ und

$$H^H(j\omega_0) + H(j\omega_0) \geq 0$$

ist. Also folgt

$$\det (1_n + H(j\omega) C_0) \neq 0 \quad \text{für alle } \omega \in \mathbf{R}.$$

Wegen

$$h_0^H + h_0 = \lim_{|\omega| \rightarrow \infty} (H^H(j\omega) + H(j\omega)) \geq 0$$

gilt auf Grund einer analogen Überlegung

$$\det (1_n + h_0 C_0) \neq 0.$$

Es bleibt noch zu zeigen, daß auch die Bedingung (iii) von Satz 27 erfüllt ist. Zunächst wollen wir aber beweisen, daß für jede $n \times n$ -Matrix F für die $F^H + F \geq 0$ und $\det (1_n + F) \neq 0$ gilt, die Ungleichung

$$\Lambda \{ (1_n + F)^{-1} F \} < 1 \tag{35}$$

erfüllt ist. Für eine beliebige quadratische $n \times n$ -Matrix M gilt

$$\Lambda^2 \{ M \} = \sup_{\substack{x \in \mathbf{C}^n \\ x \neq 0}} \frac{x^H M^H M x}{x^H x}.$$

Ferner ist $(1_n + F)^{-1} = 1_n - (1_n + F)^{-1}F$. Damit wird

$$\begin{aligned} & F^H(1_n + F)^{-1H}(1_n + F)^{-1}F \\ &= 1_n - (1_n + F)^{-1H} - (1_n + F)^{-1} + (1_n + F)^{-1H}(1_n + F)^{-1} \\ &= 1_n - (1_n + F)^{-1H}(1_n + F^H + F)(1_n + F)^{-1}. \end{aligned}$$

Da die Matrix $(1_n + F)^{-1H}(1_n + F^H + F)(1_n + F)^{-1}$ wegen des Trägheitsgesetzes quadratischer Formen positiv definit ist, folgt

$$\frac{x^H \{ (1_n + F)^{-1}F \}^H \{ (1_n + F)^{-1}F \} x}{x^H x} < 1$$

für alle $x \in \mathbb{C}^n$ mit $x \neq 0$, und damit gilt Gl. (35). Gemäß Bedingung (iii) treffen wir jetzt folgende Fallunterscheidung:

1. $C_0 = \gamma I$

Wegen $\gamma = \sum_{i=1}^n (a_i + \beta_i)$ für $i=1, \dots, n$ ist $\gamma > 0$, und es gilt folglich

$$(H(j\omega)C_0)^H + H(j\omega)C_0 = \gamma(H^H(j\omega) + H(j\omega)) \geq 0$$

für alle $\omega \in \mathbb{R}$. Somit ist

$$\begin{aligned} r_0 &= \sup_{\omega \in \mathbb{R}} \Lambda \{ (1_n + H(j\omega)C_0)^{-1} H(j\omega)T_0^{-1} \} \cdot \eta(C_0, T_0) \\ &= \sup_{\omega \in \mathbb{R}} \Lambda \{ (1_n + \gamma H(j\omega))^{-1} \gamma H(j\omega)T_0^{-1} \frac{1}{\gamma} \} \cdot \max_{i=1, \dots, n} \left(\frac{t_i^{(0)} (\beta_i - a_i)}{2} \right) \\ &\leq \sup_{\omega \in \mathbb{R}} \Lambda \{ (1_n + \gamma H(j\omega))^{-1} \gamma H(j\omega) \} \max_{i=1, \dots, n} \left(\frac{2}{(a_i + \beta_i) t_i^{(0)}} \right) \\ &\quad \cdot \max_{i=1, \dots, n} \left(\frac{t_i^{(0)} (\beta_i - a_i)}{2} \right) \end{aligned}$$

und wegen der Unabhängigkeit der Zahlen $t_i^{(0)}(\beta_i - a_i)$ von i folgt

$$r_0 \leq \sup_{\omega \in \mathbf{R}} \Lambda \{ (1_n + \gamma H(j\omega))^{-1} \gamma H(j\omega) \} \cdot \max_{i=1 \dots n} \left(\frac{\beta_i - a_i}{\beta_i + a_i} \right).$$

Auf Grund von Gl. (35) gilt mit $F = \gamma H(j\omega)$

$$\Lambda \{ (1_n + \gamma H(j\omega))^{-1} \gamma H(j\omega) \} < 1$$

für alle $\omega \in \mathbf{R}$, und damit ist

$$\sup_{\omega \in \mathbf{R}} \Lambda \{ (1_n + \gamma H(j\omega))^{-1} \gamma H(j\omega) \} \leq 1.$$

Wegen $a_i > 0$ gilt ferner

$$\max_{i=1, \dots, n} \left(\frac{\beta_i - a_i}{\beta_i + a_i} \right) < 1$$

und schließlich

$$r_0 < 1.$$

2. $T_0 = I$

Es ist hier

$$r_0 = \sup_{\omega \in \mathbf{R}} \Lambda \{ (1_n + H(j\omega) C_0)^{-1} H(j\omega) \} \cdot \max_{i=1, \dots, n} \eta_i(c_i^{(0)}).$$

Aus den Matrixumformungen

$$\begin{aligned} & (1_n + H(j\omega) C_0)^{-1} H(j\omega) \\ &= \{ C_0^{-1} (C_0 + C_0 H(j\omega) C_0) \}^{-1} H(j\omega) \\ &= \{ C_0^{-1/2} (1_n + C_0^{1/2} H(j\omega) C_0^{1/2}) C_0^{1/2} \}^{-1} H(j\omega) \\ &= C_0^{-1/2} (1_n + C_0^{1/2} H(j\omega) C_0^{1/2})^{-1} C_0^{1/2} H(j\omega) C_0^{1/2} C_0^{-1/2} \end{aligned}$$

folgt

$$\begin{aligned} & \sup_{\omega \in \mathbb{R}} \wedge \{ (1_n + H(j\omega)C_0)^{-1} H(j\omega) \} \\ &= \sup_{\omega \in \mathbb{R}} \wedge \{ C_0^{-1/2} (1_n + C_0^{1/2} H(j\omega) C_0^{1/2})^{-1} C_0^{1/2} H(j\omega) C_0^{1/2} C_0^{-1/2} \} \\ &\leq \sup_{\omega \in \mathbb{R}} \wedge \{ (1_n + C_0^{1/2} H(j\omega) C_0^{1/2})^{-1} C_0^{1/2} H(j\omega) C_0^{1/2} \} \cdot \left[\max_{i=1, \dots, n} \left(\frac{1}{\sqrt{c_i(0)}} \right) \right]^2 \end{aligned}$$

Also gilt

$$\begin{aligned} r_0 &\leq \sup_{\omega \in \mathbb{R}} \{ (1_n + C_0^{1/2} H(j\omega) C_0^{1/2})^{-1} C_0^{1/2} H(j\omega) C_0^{1/2} \} \\ &\quad \cdot \max_{i=1, \dots, n} \left(\frac{\beta_i - a_i}{\beta_i + a_i} \right). \end{aligned}$$

Wegen

$$\begin{aligned} (C_0^{1/2} H(j\omega) C_0^{1/2})^H + C_0^{1/2} H(j\omega) C_0^{1/2} &= C_0^{1/2} (H^H(j\omega) + H(j\omega)) C_0^{1/2} \\ &\geq 0 \end{aligned}$$

ist mit Gl. (35)

$$\sup_{\omega \in \mathbb{R}} \wedge \{ (1_n + C_0^{1/2} H(j\omega) C_0^{1/2})^{-1} C_0^{1/2} H(j\omega) C_0^{1/2} \} \leq 1.$$

Weiter ist

$$\max_{i=1, \dots, n} \left(\frac{\beta_i - a_i}{\beta_i + a_i} \right) < 1,$$

und somit gilt schließlich

$$r_0 < 1.$$

q.e.d.

Analog zu Satz 29 gilt für die Operatorgl. (2):

Satz 30: Es seien die Voraussetzungen von Satz 28 gegeben. Falls die Bedingungen

- (i) $a_i > 0$ für $i=1, \dots, n$,
- (ii) $H^H(j\omega) + H(j\omega) \geq 0$ für alle $\omega \in \mathbf{R}$ und
- (iii) $\mathbf{T}_0 = \mathbf{I}$ oder $\mathbf{C}_0 = \gamma \mathbf{I}$ mit $\gamma \in \mathbf{R}$

erfüllt sind, gelten auch die Bedingungen (i), (ii) und (iii) von Satz 28.

Beweis:

Der Beweis erfolgt analog zum Beweis von Satz 29.

Bemerkung 1:

Falls der nichtlineare Systemanteil passiv ist, kann die Bedingung $\mathbf{T}_0 = \mathbf{I}$ oder $\mathbf{C}_0 = \gamma \mathbf{I}$ durch eine geeignete Wahl der Vektoren α und β unter Beibehaltung der Passivität von $\mathbf{N}$ immer erfüllt werden.

Bemerkung 2:

Sandberg hat in [14] nachgewiesen, daß unter den Voraussetzungen und Bedingungen von Satz 29

$$\gamma_0 \in \mathbf{P}_1 \text{ und } \hat{f}(\gamma_0 \mathbf{I}, \mathbf{I}) < 1$$

gilt. (Dabei werden α , β , $\mathbf{C}$ und $\mathbf{T}$ im Sinne von Abschnitt II.2.4.4 so gewählt, daß Sandbergs Theorie anwendbar wird.) Die Aussagen von Satz 29 können somit als Konsequenz der Sätze 20 und 22 aufgefaßt werden und sie hätten deshalb nicht explizit bewiesen werden müssen. Das Analoge gilt für Satz 30.

2.5.3.2 Die Operatorgleichung $e = f + \mathbf{H}\mathbf{N}f$ bei $n = 1$

Im Falle $n = 1$ (das System enthält dann nur ein nichtlineares Element) lauten die Bedingungen (i), (ii) und (iii) von Satz 27 auf Grund der Bemerkung von Abschnitt II.2.5.3 folgendermaßen:

- (i) $1/2(a + \beta)H(j\omega) = -1$ für alle $\omega \in \mathbf{R}$,
- (ii) $1/2(a + \beta)h_0 = -1$,
- (iii) $1/2(\beta - a) \cdot \sup_{\omega \in \mathbf{R}} \left| \frac{H(j\omega)}{1 + 1/2(a + \beta)H(j\omega)} \right| < 1$.

Diese drei Bedingungen schränken bei vorgegebenem a und β den zulässigen Frequenzgang

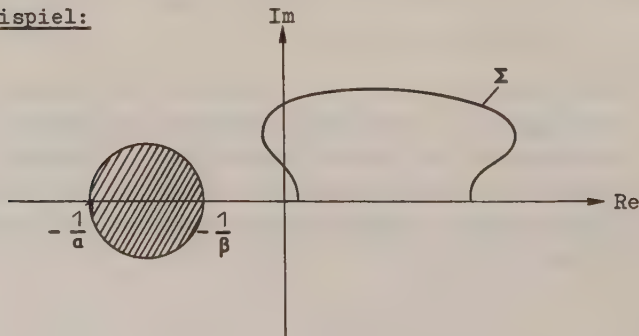
$$\Sigma = \{H(j\omega) \mid \omega \in \mathbf{R}\} \cup \{h_0\}$$

des linearen Systemanteils ein. Dabei ergibt sich für das Erfülltsein dieser Bedingungen ein interessantes geometrisches Kriterium, das anhand einfacher Umformung ermittelt werden kann. Wir unterscheiden drei Fälle:

1. $a > 0$

Falls der Frequenzgang Σ außerhalb des Kreises mit dem Mittelpunkt $-1/2(1/a + 1/\beta)$ und dem Radius $1/2(1/a - 1/\beta)$ liegt, sind die Bedingungen (i), (ii) und (iii) erfüllt.

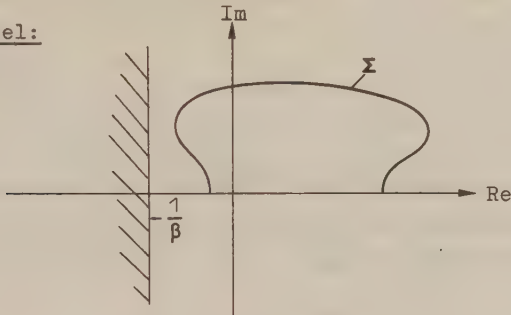
Beispiel:



2. $a = 0$

Falls der Frequenzgang Σ rechts der Parallelen zur imaginären Achse, die durch den Punkt $-1/\beta$ verläuft, liegt, sind die Bedingungen (i), (ii) und (iii) erfüllt.

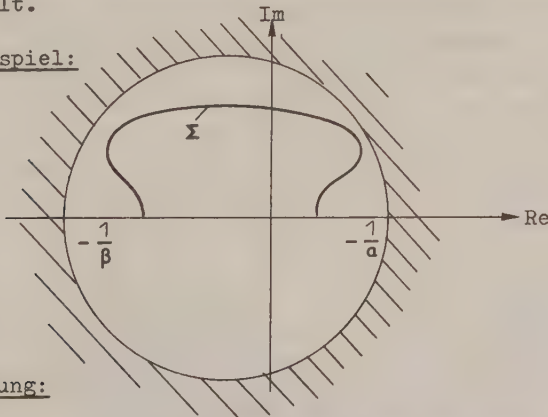
Beispiel:



3. $a < 0, \beta > 0$

Falls der Frequenzgang Σ innerhalb des Kreises mit dem Mittelpunkt $-1/2(1/a + 1/\beta)$ und dem Radius $1/2(1/\beta - 1/a)$ verläuft, sind die Bedingungen (i), (ii) und (iii) erfüllt.

Beispiel:



Bemerkung:

Der Fall $a < 0$ und $\beta < 0$ kann durch eine einfache Transformation in den Fall 1 übergeführt werden.

2.5.3.3 Die Operatorgleichung $e = \mathbf{N}f + \mathbf{H}f$ bei $n = 1$

Im Fall $n = 1$ (das System enthält dann nur ein nichtlineares Element) lauten die Bedingungen (i), (ii) und (iii) von Satz 28 auf Grund der Bemerkung von Abschnitt II.2.5.3 folgendermaßen:

- (i) $H(j\omega) = -1/2(\alpha + \beta)$ für alle $\omega \in \mathbf{R}$,
- (ii) $h_0 = -1/2(\alpha + \beta)$,
- (iii) $1/2(\beta - \alpha) \cdot \sup_{\omega \in \mathbf{R}} \left| \frac{1}{1/2(\alpha + \beta) + H(j\omega)} \right| < 1$.

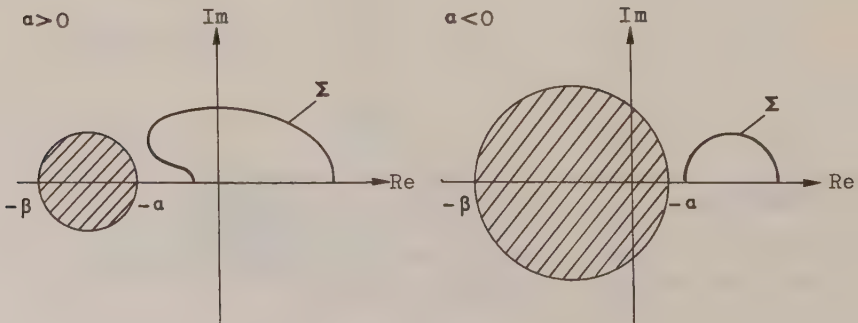
Diese drei Bedingungen schränken bei vorgegebenem α und β den zulässigen Frequenzgang

$$\Sigma = \{H(j\omega) | \omega \in \mathbf{R}^+\} \cup \{h_0\}$$

des linearen Systemanteils ein. Dabei ergibt sich für das Erfülltsein dieser Bedingungen ein interessantes geometrisches Kriterium, das anhand einfacher Umformungen ermittelt werden kann:

Falls der Frequenzgang Σ außerhalb des Kreises mit dem Mittelpunkt $-1/2(\alpha + \beta)$ und dem Radius $1/2(\beta - \alpha)$ verläuft, sind die Bedingungen (i), (ii) und (iii) erfüllt.

Beispiele:



3. Allgemeine Systembetrachtungen

In diesem Kapitel wollen wir einige Fragen behandeln, die in engem Zusammenhang mit der in Abschnitt II.2 behandelten Systemtheorie stehen. Insbesondere soll dabei gezeigt werden, daß diese Systemtheorie auch auf komplexe Systemstrukturen angewandt werden kann. Ferner wollen wir die Möglichkeit der Approximation von nichtlinearen Rückkopplungssystemen durch nichtlineare Geradeaussysteme diskutieren. Abschließend wollen wir andeuten, wie die Theorie von Abschnitt II.2 sinngemäß auch auf solche Systemstrukturen angewandt werden kann, die nicht auf die Standardformen von Abschnitt II.2 reduziert werden können.

3.1 Komplexe Systemstrukturen

Die in der Praxis auftretenden Systemblockschaltbilder sind im allgemeinen viel komplizierter als die in Abb. (1) bzw. Abb. (2) von Kapitel II.2 dargestellten. Anhand des folgenden Beispiels (Abb. (12)) wird gezeigt, wie man in vielen Fällen ein komplexes Rückkopplungssystem sukzessive in ein äquivalentes Rückkopplungssystem unseres Standardtyps umwandeln kann. Dabei wird vorausgesetzt, daß alle vorkommenden Operatoren den selben Definitions- und Bildbereich besitzen. Die G_i sind dabei lineare Operatoren, während N ein nichtlinearer Operator ist.

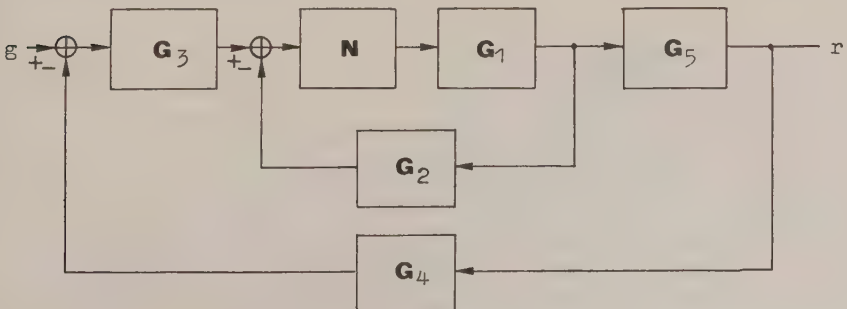


Abb. (12 a)

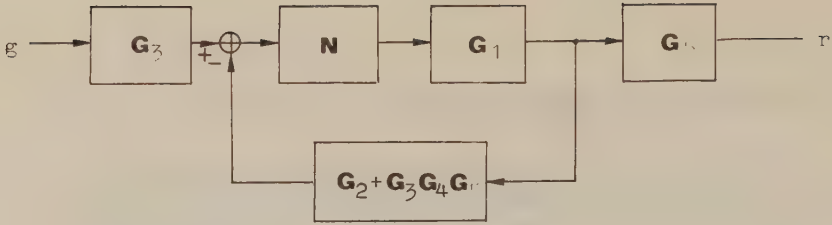


Abb. (12 b)

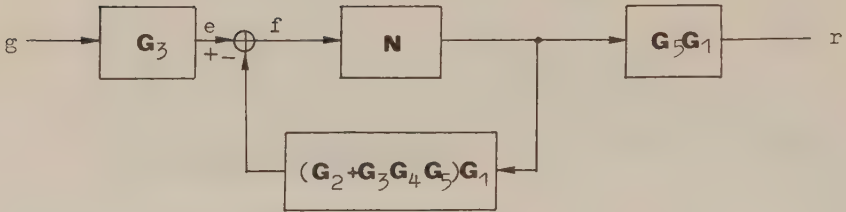


Abb. (12 c)

Mit den Abkürzungen

$$H = (G_2 + G_3G_4G_5)G_1 \quad \text{und} \quad e = G_3g$$

kommen wir zum Blockschaltbild Abb. (13), das wir anhand des Kapitels II.2 untersuchen können. Die Systemreaktion r ergibt sich schließlich zu

$$r = G_5G_1Nf.$$

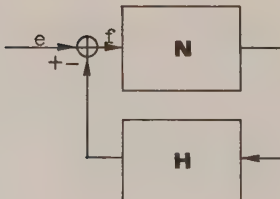


Abb. (13)

3.2 Anfangsbedingungen

Untersucht man ein vorgegebenes Rückkopplungssystem, so wird dieses am Anfang des Beobachtungszeitraums bestimmten Anfangsbedingungen genügen, die man häufig dadurch beschreiben kann, daß man bei den linearen Systemblöcken $\mathbf{G}_i$ zusätzliche Erregungen g_i anbringt (siehe Abb. (14)).



Abb. (14)

Im folgenden Beispiel (Abb. (15)) wird gezeigt, wie man durch sukzessive Umwandlungen aus einem Blockschaltbilde mit mehreren Erregungen ein äquivalentes Blockschaltbild unseres Standardtyps bekommt. Wie im vorhergehenden Abschnitt haben die linearen Operatoren $\mathbf{G}_i$ und der nichtlineare Operator $\mathbf{N}$ den selben Definitions- und Bildbereich. Alle Erregungen g_i seien ebenfalls aus diesem Raum.

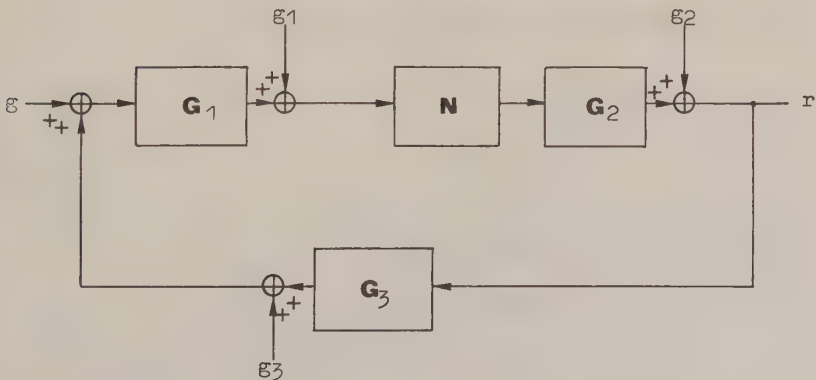


Abb. (15 a)

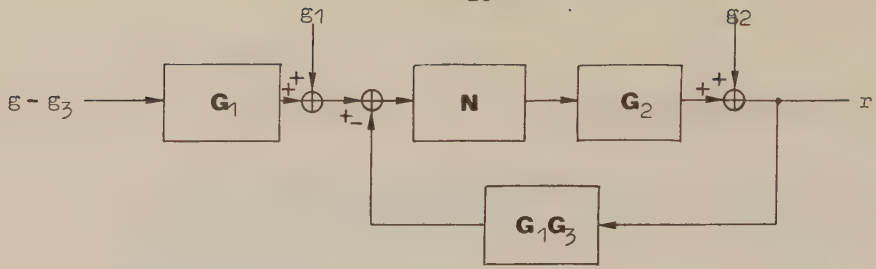


Abb. (15 b)

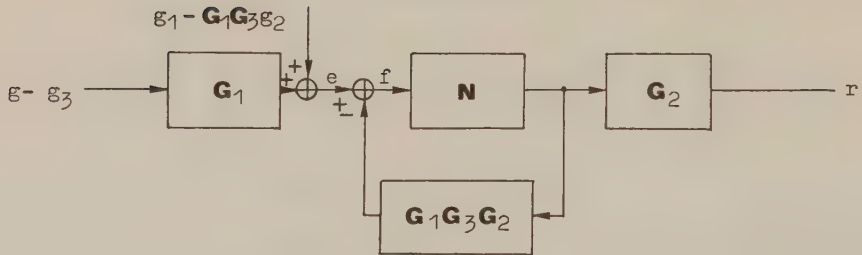


Abb. (15 c)

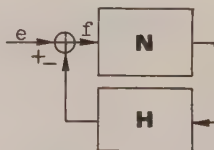
Mit den Abkürzungen

$$e = s_1 + G_1(s - s_3) - G_1G_3s_2$$

und

$$H = G_1G_2G_3$$

erhalten wir unser Standardblockschaltbild



Die Systemreaktion r ergibt sich zu

$$r = G_2 N f.$$

3.3 Systemapproximation

Gegeben sei die Operatorgleichung

$$e = f + \mathbf{H} \mathbf{N} f, \quad (36)$$

und es seien alle Voraussetzungen und Bedingungen von Satz 2 erfüllt. Dann gilt gemäß Aussage dieses Satzes

$$f = \lim_{\mu \rightarrow \infty} f_{\mu}$$

mit

$$f_{\mu+1} = -(\mathbf{I} + \mathbf{H} \mathbf{C}_0)^{-1} \mathbf{H} (\mathbf{N} - \mathbf{C}_0) f_{\mu} + (\mathbf{I} + \mathbf{H} \mathbf{C}_0)^{-1} e$$

und

$$f_0 \in \mathbf{B}^n.$$

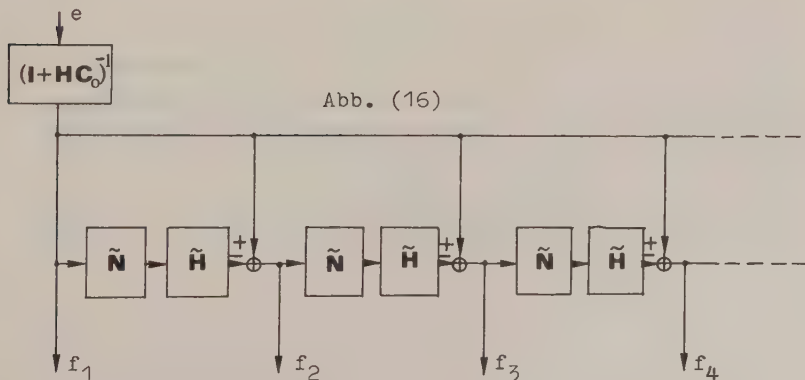
Weiter ist

$$\|f - f_{\mu}\| \leq \frac{r_0^{\mu}}{1 - r_0} \cdot \|f_1 - f_0\|. \quad (37)$$

Falls nun $r_0 < 1$ hinreichend klein ist, wird man f_{μ} für geeignet gewähltes μ als gute Approximation von f auffassen können. Wählt man nun f_0 gleich den Nullelement, kann man sich die Approximationen $f_1, f_2, \dots, f_{\mu}$ durch folgendes Blockschaltbild entstanden denken, wobei wir zur Abkürzung

$$\tilde{\mathbf{H}} = (\mathbf{I} + \mathbf{H} \mathbf{C}_0)^{-1} \mathbf{H} \quad \text{und} \quad \tilde{\mathbf{N}} = \mathbf{N} - \mathbf{C}_0$$

setzen:



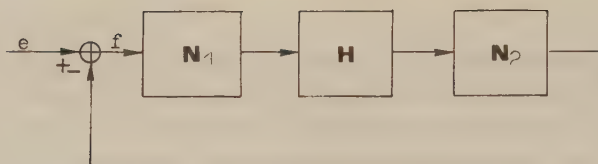


Abb. (17)

Die Transformation von Gl. (40) erfolgt auf folgende Weise:

$$\begin{aligned}
 e &= f + (\mathbf{N}_2 - \mathbf{C}_2)\mathbf{H}\mathbf{N}_1 f + \mathbf{C}_2\mathbf{H}\mathbf{N}_1 f \\
 &= f + (\mathbf{N}_2 - \mathbf{C}_2)\mathbf{H}\mathbf{N}_1 f + \mathbf{C}_2\mathbf{H}(\mathbf{N}_1 - \mathbf{C}_1)f + \mathbf{C}_2\mathbf{H}\mathbf{C}_1 f.
 \end{aligned} \tag{40}$$

Wir wollen jetzt voraussetzen, daß der durch die Operatoren $\mathbf{C}_1$ und $\mathbf{C}_2$ linearisierte Systemoperator $(\mathbf{I} + \mathbf{C}_2\mathbf{H}\mathbf{C}_1)^{-1}$ existiert. Dann folgt aus Gl. (41)

$$\begin{aligned}
 f &= -(\mathbf{I} + \mathbf{C}_2\mathbf{H}\mathbf{C}_1)^{-1} \{ (\mathbf{N}_2 - \mathbf{C}_2)\mathbf{H}\mathbf{N}_1 + \mathbf{C}_2\mathbf{H}(\mathbf{N}_1 - \mathbf{C}_1) \} f \\
 &\quad + (\mathbf{I} + \mathbf{C}_2\mathbf{H}\mathbf{C}_1)^{-1} e
 \end{aligned}$$

oder kurz

$$f = \mathbf{R}f + \mathbf{L}e.$$

Falls jetzt noch $g_0(\mathbf{R}) < 1$ ist, existiert der Systemoperator

$$(\mathbf{I} + \mathbf{N}_2\mathbf{H}\mathbf{N}_1)^{-1}.$$

Dies ist z.B. sicher erfüllt, wenn

$$\|(\mathbf{I} + \mathbf{C}_2\mathbf{H}\mathbf{C}_1)^{-1}\| \cdot \{g_0(\mathbf{N}_2 - \mathbf{C}_2) \cdot g_0(\mathbf{N}_1) + g_0(\mathbf{C}_2) \cdot g_0(\mathbf{N}_1 - \mathbf{C}_1)\} \cdot \|\mathbf{H}\| < 1$$

gilt.

- (i) Gl. (38) und Gl. (39) sind äquivalent, d.h. der Operator $(\mathbf{I} + \mathbf{HC})^{-1}$ existiert.
- (ii) $g_0(\mathbf{R}) < 1$.

Es muß also der linearisierte Systemoperator existieren und die nichtlineare "Störung" in gewissem Sinne klein sein.

Die Lösbarkeit der Operatorgleichung $e = \mathbf{N}f + \mathbf{H}f$ wurde in Abschnitt II.2 auf völlig analoge Weise untersucht.

Da in der Praxis häufig kompliziertere Operatorgleichungen z.B. mit mehreren nichtlinearen Operatoren $\mathbf{N}_i$ auftreten, wollen wir auch für diese Fälle eine den obigen Überlegungen analoge Vorgehensweise vorschlagen:

- (1) Transformation der Operatorgleichung durch geeignete Linearisierung der nichtlinearen Operatoren $\mathbf{N}_i$ durch die linearen Operatoren $\mathbf{C}_i$ in eine äquivalente Gleichung der Form $f = \mathbf{R}f + \mathbf{L}e$. Dabei enthält $\mathbf{R}$ alle nichtlinearen Restoperatoren $(\mathbf{N}_i - \mathbf{C}_i)$, und $\mathbf{L}$ ist der Operator des linearisierten Gesamtsystems.
- (2) Falls $g_0(\mathbf{R}) < 1$ ist, kann auf die transformierte Gleichung der Banachsche Fixpunktsatz angewandt werden, und damit existiert eine eindeutige Lösung f , die mittels eines Iterationsverfahrens berechnet werden kann.

Dies soll anhand eines Beispiels demonstriert werden:

Gegeben sei ein System gemäß Abb. (17), das der Gleichung

$$e = f + \mathbf{N}_2 \mathbf{H} \mathbf{N}_1 f \quad (40)$$

genügt.

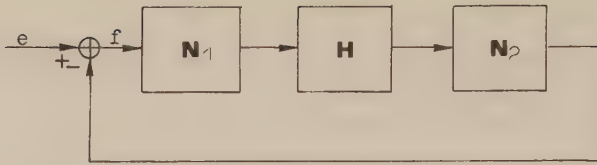


Abb. (17)

Die Transformation von Gl. (40) erfolgt auf folgende Weise:

$$\begin{aligned}
 e &= f + (\mathbf{N}_2 - \mathbf{C}_2)\mathbf{H}\mathbf{N}_1 f + \mathbf{C}_2\mathbf{H}\mathbf{N}_1 f \\
 &= f + (\mathbf{N}_2 - \mathbf{C}_2)\mathbf{H}\mathbf{N}_1 f + \mathbf{C}_2\mathbf{H}(\mathbf{N}_1 - \mathbf{C}_1)f + \mathbf{C}_2\mathbf{H}\mathbf{C}_1 f.
 \end{aligned} \tag{40}$$

Wir wollen jetzt voraussetzen, daß der durch die Operatoren $\mathbf{C}_1$ und $\mathbf{C}_2$ linearisierte Systemoperator $(\mathbf{I} + \mathbf{C}_2\mathbf{H}\mathbf{C}_1)^{-1}$ existiert. Dann folgt aus Gl. (41)

$$\begin{aligned}
 f &= -(\mathbf{I} + \mathbf{C}_2\mathbf{H}\mathbf{C}_1)^{-1} \{ (\mathbf{N}_2 - \mathbf{C}_2)\mathbf{H}\mathbf{N}_1 + \mathbf{C}_2\mathbf{H}(\mathbf{N}_1 - \mathbf{C}_1) \} f \\
 &\quad + (\mathbf{I} + \mathbf{C}_2\mathbf{H}\mathbf{C}_1)^{-1} e
 \end{aligned}$$

oder kurz

$$f = \mathbf{R}f + \mathbf{L}e.$$

Falls jetzt noch $g_0(\mathbf{R}) < 1$ ist, existiert der Systemoperator

$$(\mathbf{I} + \mathbf{N}_2\mathbf{H}\mathbf{N}_1)^{-1}.$$

Dies ist z.B. sicher erfüllt, wenn

$$\|(\mathbf{I} + \mathbf{C}_2\mathbf{H}\mathbf{C}_1)^{-1}\| \cdot \{g_0(\mathbf{N}_2 - \mathbf{C}_2) \cdot g_0(\mathbf{N}_1) + g_0(\mathbf{C}_2) \cdot g_0(\mathbf{N}_1 - \mathbf{C}_1)\} \cdot \|\mathbf{H}\| < 1$$

gilt.

III. ANALYSE EINER KLASSE NICHTLINEARER ZEITVARIANTER ELEKTRISCHER NETZWERKE

1. Einleitung

In diesem Kapitel werden wir die in Abschnitt II.2.5 beschriebene Theorie auf elektrische Netzwerke anwenden. Wie im folgenden gezeigt wird, können elektrische Netzwerke, die außer linearen Elementen nichtlineare zeitvariante Widerstände und nichtlineare zeitvariante gesteuerte Quellen enthalten, durch die Operatorgleichungen

$$e = f + H N f \quad (1)$$

oder

$$e = N f + H f \quad (2)$$

beschrieben werden. Dabei bedeutet $e \in L_{2r}^n(-\infty, \infty)$ die Vektorfunktion der Ersatzspannungs- bzw. Ersatzstromquellen (damit können auch Anfangsbedingungen dargestellt werden) und f repräsentiert die Vektorfunktion der Ströme bzw. Spannungen in bzw. an den nichtlinearen Elementen. Bemerkenswert ist, daß der lineare Netzwerkanteil, der durch den Faltungsoperator H dargestellt wird, nicht notwendig passiv, reziprok und konzentriert sein muß. Anhand der Sätze II.27 bzw. II.28 sind wir unter gewissen Bedingungen in der Lage, sicherzustellen, daß für jede quadratisch integrierbare Erregung des Netzwerks ein eindeutiges $f \in L_{2r}^n(-\infty, \infty)$ existiert, das die Gl. (1) bzw. (2) befriedigt. Des weiteren können wir die gesuchte Lösung f mit einem Iterationsverfahren berechnen. In Abschnitt III.3.4 behandeln wir schließlich die Gleichstromanalyse einer Klasse elektrischer Netzwerke.

2. Vereinbarungen

Der den Abschnitten III.3.1, III.3.2 und III.3.3 zugrundegelegte Banachraum $\mathbf{B}^n$ sei der Raum $\mathbf{L}_{2r}^n(-\infty, \infty)$, und wir wollen dort wie in Abschnitt II.2.5 annehmen, daß $\mathbf{N} \in \Psi(\alpha, \beta)$ ist. In Abschnitt III.3.4 wird $\mathbf{N}$ gesondert definiert. Dem nichtlinearen Operator $\mathbf{N}$ sind die Matrizen

$$\mathbf{C}_0 = \text{diag} (c_1^{(0)}, c_2^{(0)}, \dots, c_n^{(0)})$$

und

$$\mathbf{T}_0 = \text{diag} (t_1^{(0)}, t_2^{(0)}, \dots, t_n^{(0)})$$

zugeordnet. Dabei ist $c_i^{(0)} = (\alpha_i + \beta_i)/2$ für $i = 1, \dots, n$. Die reellen $t_i^{(0)}$ sind durch $t_i^{(0)}(\alpha_i - \beta_i) = t_j^{(0)}(\alpha_j - \beta_j)$ für $i, j = 1, \dots, n$ und $t_i^{(0)} > 0$ für $i = 1, \dots, n$ definiert (siehe Abschnitt II.2.4.3). Schließlich ist

$$\eta(\mathbf{C}_0, \mathbf{T}_0) = \max_{i=1, \dots, n} \left(\frac{t_i^{(0)}(\beta_i - \alpha_i)}{2} \right).$$

Falls ein linearer Operator $\mathbf{H}$ ein Faltungsoperator gemäß Abschnitt I.4.4 ist, schreiben wir $\mathbf{H} \in \mathbf{F}$ und verstehen unter $\mathbf{H}(j\omega)$ die diesem Faltungsoperator zugeordnete Fourier-Transformierte.

3. Einige Klassen nichtlinearer zeitvarianter elektrischer Netzwerke

3.1 Elektrische Netzwerke mit nichtlinearen zeitvarianten Widerständen

Gegeben sei ein elektrisches Netzwerk, das lineare zeitinvariante (aber nicht notwendig konzentrierte, passive und

reziproke) Elemente und n nichtlineare zeitvariante Widerstände bzw. Leitwerte enthält. Die Energiespeicher seien zum Zeitpunkt $t = -\infty$ ungeladen. Erregt werde dieses Netzwerk durch Strom- und Spannungsquellen mit quadratisch integrierbarem zeitlichen Verlauf. Anfangsbedingungen der Energiespeicher können durch diese Quellen berücksichtigt werden [23]. Wir haben also die in Abb. (1) dargestellte Situation. Dabei bedeutet:

$e_0 \in L_{2r}^r(-\infty, \infty)$ Vektorfunktion der r Spannungsquellen,

$i_0 \in L_{2r}^s(-\infty, \infty)$ Vektorfunktion der s Stromquellen,

i_1, u_1 Vektorfunktion der u Ströme bzw. Spannungen in bzw. an den nichtlinearen zeitvarianten Widerständen,

i_2, u_2 Vektorfunktion der v Ströme bzw. Spannungen in bzw. an den nichtlinearen zeitvarianten Leitwerten,

i, u Vektorfunktion der gesuchten w Ströme und t Spannungen des Netzwerks.

Offenbar ist $n = u + v$ und ferner gelte $m = r + s$ und $d = w + t$.

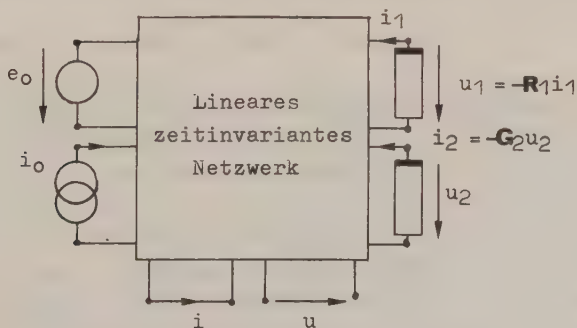


Abb. (1)

Wir nehmen an, daß im linearen Netzwerkanteil das Thevenin-Norton-Theorem in der Form

$$\begin{bmatrix} i_1 \\ u_2 \end{bmatrix} = \mathbf{H} \begin{bmatrix} u_1 \\ i_2 \end{bmatrix} + \mathbf{U} \begin{bmatrix} e_0 \\ i_0 \end{bmatrix} \quad (3)$$

für alle

$$\begin{bmatrix} u_1 \\ i_2 \end{bmatrix} \in \mathbf{L}_{2r}^n(-\infty, \infty) \text{ und } \begin{bmatrix} e_0 \\ i_0 \end{bmatrix} \in \mathbf{L}_{2r}^m(-\infty, \infty)$$

gilt, wobei die linearen Operatoren $\mathbf{H}$ und $\mathbf{U}$ Faltungsoperatoren seien, d.h. es ist $\mathbf{H}, \mathbf{U} \in \mathbf{F}$. Gl. (3) beschreibt das Verhalten des linearen Netzwerkanteils bezüglich der n Tore vollständig. Um eine Gleichung für das gesamte Netzwerk formulieren zu können, benötigen wir noch die Eigenschaften des nichtlinearen zeitvarianten Netzwerkanteils:

Die u nichtlinearen Widerstände und v Leitwerte seien durch

$$u_{1i}(t) = -r_i(i_{1i}(t), t)$$

und

$$i_{2i}(t) = -g_i(u_{2i}(t), t)$$

für $i = 1, \dots, n$ oder kurz

$$u_1 = -\mathbf{R}_1 i_1 \text{ und } i_2 = -\mathbf{G}_2 u_2$$

definiert. Um die Theorie von Kapitel II. anwenden zu können, nehmen wir an, daß die nichtlinearen Funktionen $r_i(x, t)$ und $g_i(x, t)$ die Bedingungen von Abschnitt II.2.3 erfüllen, so daß mit

$$\begin{bmatrix} u_1 \\ i_2 \end{bmatrix} = - \begin{bmatrix} \mathbf{R}_1 i_1 \\ \mathbf{G}_2 u_2 \end{bmatrix} = -\mathbf{N} \begin{bmatrix} i_1 \\ u_2 \end{bmatrix} \quad (4)$$

schließlich $\mathbf{N} \in \Psi(a, \beta)$ gilt.

Aus Gl. (3) und Gl. (4) folgt schließlich die Gleichung

$$\mathbf{U} \begin{bmatrix} e_0 \\ i_0 \end{bmatrix} = \begin{bmatrix} i_1 \\ u_2 \end{bmatrix} + \mathbf{H}\mathbf{N} \begin{bmatrix} i_1 \\ u_2 \end{bmatrix}$$

oder kurz

$$\mathbf{e} = \mathbf{f} + \mathbf{H}\mathbf{N}\mathbf{f} \quad (5)$$

für die gesuchte Vektorfunktion

$$\mathbf{f} = \begin{bmatrix} i_1 \\ u_2 \end{bmatrix}.$$

Wegen $\mathbf{U} \in \mathbf{F}$ ist

$$\mathbf{e} = \mathbf{U} \begin{bmatrix} e_0 \\ i_0 \end{bmatrix} \in \mathbf{L}_{2r}^n(-\infty, \infty),$$

und damit sind bis auf die Bedingungen (i), (ii) und (iii) die Voraussetzungen des Satzes II.27 bezüglich Gl. (5) erfüllt. Um die Existenz eines eindeutigen

$$\mathbf{f} = \begin{bmatrix} i_1 \\ u_2 \end{bmatrix} \in \mathbf{L}_{2r}^n(-\infty, \infty)$$

zu sichern, müssen also noch die Bedingungen

(i) $\det(\mathbf{1}_n + \mathbf{H}(j\omega)\mathbf{C}_0) \neq 0$ für alle $\omega \in \mathbf{R}$,

(ii) $\det(\mathbf{1}_n + \mathbf{h}_0\mathbf{C}_0) \neq 0$ und

(iii) $r_0 = \sup_{\omega \in \mathbf{R}} \wedge \{ (\mathbf{1}_n + \mathbf{H}(j\omega)\mathbf{C}_0)^{-1} \mathbf{H}(j\omega)\mathbf{T}_0^{-1} \} \eta(\mathbf{C}_0, \mathbf{T}_0) < 1$

erfüllt sein, und dann ist es auch möglich, $\mathbf{f}$ anhand eines Iterationsverfahrens zu berechnen. Die gesuchten Ströme i und Spannungen u lassen sich schließlich mit bekannten Methoden aus der linearen Netzwerktheorie[23] berechnen (Kompensationstheorem), und es ist

$$r = \begin{bmatrix} u \\ i \end{bmatrix} = \mathbf{V} \begin{bmatrix} u_1 \\ i_2 \end{bmatrix} + \mathbf{S} \begin{bmatrix} e_o \\ i_o \end{bmatrix}.$$

Falls $\mathbf{V}, \mathbf{S} \in \mathbf{F}$ ist, gilt

$$r = \begin{bmatrix} u \\ i \end{bmatrix} \in \mathbf{L}_{2r}^d(-\infty, \infty).$$

Das gesamte Netzwerkverhalten kann also bezüglich der für uns interessanten Größen durch das Systemblockschaltbild Abb. (2) dargestellt werden.

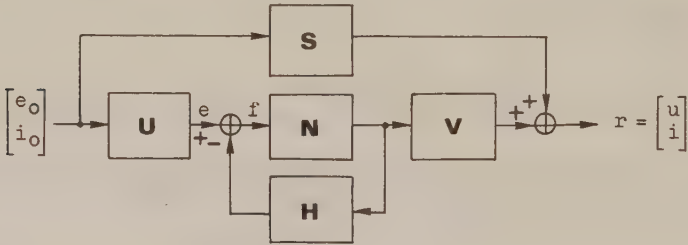


Abb. (2)

Bemerkung 1:

Falls für den linearen Netzwerkanteil

$$\begin{bmatrix} u_1 \\ i_2 \end{bmatrix} = \mathbf{H} \begin{bmatrix} i_1 \\ u_2 \end{bmatrix} + \mathbf{U} \begin{bmatrix} e_o \\ i_o \end{bmatrix}$$

mit $\mathbf{H}, \mathbf{U} \in \mathbf{F}$ gilt, folgt anhand Gl. (4)

$$-\mathbf{U} \begin{bmatrix} e_o \\ i_o \end{bmatrix} = \mathbf{N} \begin{bmatrix} i_1 \\ u_2 \end{bmatrix} + \mathbf{H} \begin{bmatrix} i_1 \\ u_2 \end{bmatrix}$$

oder kurz

$$e = \mathbf{N}f + \mathbf{H}f.$$

Gleichungen dieser Art können mit Satz II.28 untersucht werden.

Bemerkung 2:

Die Bedingungen (i), (ii) und (iii) sind nach Satz 29 sicher dann erfüllt, wenn

- (i) $T_0 = I$ oder $C_0 = \gamma I$ mit $\gamma \in \mathbb{R}$,
- (ii) $a_i > 0$ für $i=1, \dots, n$ und
- (iii) $H^H(j\omega) + H(j\omega) \geq 0$ für alle $\omega \in \mathbb{R}$

gilt. Diese Bedingungen können durch geeignete Wahl der $a_i > 0$ immer dann erfüllt werden, wenn der lineare Netzwerkanteil passiv ist (das ist dann der Fall, wenn alle linearen Netzwerkelemente passiv sind) und die nichtlinearen Widerstände bzw. Leitwerte zu jedem Zeitpunkt eine streng monoton steigende und lipschitzstetige Kennlinie aufweisen.

Bemerkung 3:

Für den Fall, daß alle linearen Netzwerkelemente konzentriert sind, können die Bedingungen (i) und (ii) durch die Bedingung

- (i) Die Matrix $\hat{H}(p) = (I_n + H(p)C_0)^{-1}$ hat auf der imaginären Achse einschließlich des Punktes $p = \infty$ keine Pole

ersetzt werden. Dabei bedeutet $H(p)$ die analytische Fortsetzung von $H(j\omega)$ auf einen Streifen der p -Ebene (Laplace-Transformation). Alle Elemente der Matrizen $H(p)$ und $\hat{H}(p)$ sind gebrochen rationale Funktionen des komplexen Frequenzparameters p .

3.1.1 RLCUT-Netzwerke mit passiven nichtlinearen zeitvarianten Widerständen

Es sei ein Netzwerk gegeben, das aus einer endlichen Anzahl von linearen, passiven und zeitinvarianten Widerständen, Kondensatoren, Induktivitäten, idealen Übertragern und Gyrotoren besteht und ferner n nach Abschnitt II.2.5.3.1 passive nichtlineare zeitvariante Widerstände enthält. Diese Nichtlinearitäten haben also zu jedem Zeitpunkt eine streng monoton steigende und lipschitzstetige Stromspannungskennlinie, und sie können wahlweise als Widerstände oder Leitwerte definiert werden. (Auf Grund der vorausgesetzten strengen Monotonie und Lipschitzstetigkeit existieren immer die inversen Funktionen, die ebenfalls zu jedem Zeitpunkt streng monoton steigend und lipschitzstetig sind, und wir wollen voraussetzen, daß auch sie die Eigenschaften gemäß Abschnitt II.2.3 aufweisen.) Das Netzwerk werde durch Strom- und Spannungsquellen erregt, und alle Energiespeicher seien zum Zeitpunkt $t = -\infty$ ungeladen.

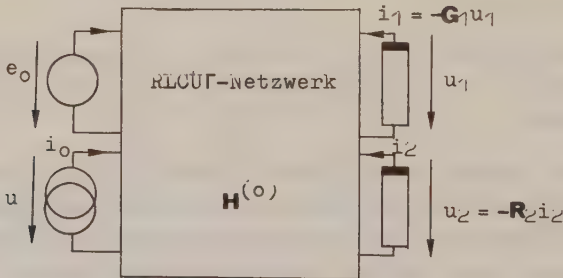


Abb. (3)

Wir haben also die in Abb. (3) gezeigte Situation und dabei bedeutet:

$e_o \in L_{2r}^r(-\infty, \infty)$ Vektorfunktion der r Spannungsquellen,

$i_o \in L_{2r}^s(-\infty, \infty)$ Vektorfunktion der s Stromquellen,

i_1, u_1 Vektorfunktion der u Ströme bzw. Spannungen in bzw. an den nichtlinearen zeitvarianten Leitwerten,

i_2, u_2 Vektorfunktion der v Ströme bzw. Spannungen in bzw. an den nichtlinearen zeitvarianten Widerständen.

Es ist $n = u + v$, und es sei $m = r + s$. Durch e_o und i_o können sowohl die Erregungen als auch die Anfangszustände des Netzwerks beschrieben werden [23]. Die nichtlinearen Elemente werden gemäß Abb. (3) so aufgeteilt, daß alle Komponenten von u_1 und i_2 als voneinander unabhängige Erregungen gewählt werden können. (Nach [2,6] existiert für jedes RLCU n -Tor wenigstens eine Hybridmatrix.) Wir nehmen ferner an, daß alle Komponenten von u_1, i_2, e_o und i_o als voneinander linear unabhängige Erregungen gewählt werden können, und nennen den linearen Netzwerkanteil erregungsqualifiziert, da diese Forderung ja nur eine Einschränkung bezüglich der Erregungen darstellt. (Bei praktisch interessanten Fällen wird der lineare Netzwerkanteil im allgemeinen erregungsqualifiziert sein.) Somit existiert folgende Hybridmatrixverknüpfung:

$$\begin{bmatrix} I_1(p) \\ U_2(p) \\ I(p) \\ U(p) \end{bmatrix} = H(o)(p) \begin{bmatrix} U_1(p) \\ I_2(p) \\ E_o(p) \\ I_o(p) \end{bmatrix} = \begin{bmatrix} H_{11}^{(o)}(p) & H_{12}^{(o)}(p) \\ H_{21}(p) & H_{22}^{(o)}(p) \end{bmatrix} \begin{bmatrix} U_1(p) \\ I_2(p) \\ E_o(p) \\ I_o(p) \end{bmatrix}, \quad (6)$$

wobei sämtliche Ströme und Spannungen die zweiseitigen Laplace-Transformierten der entsprechenden Größen im Zeitbereich sind. Dabei sind die Matrizen $H^{(o)}(p)$, $H_{11}^{(o)}(p)$ und $H_{22}^{(o)}(p)$ positiv, reell und rational [6]. Gl. (6) lautet im Zeitbereich und in Operatorschreibweise:

$$\begin{bmatrix} i_1 \\ u_2 \\ i \\ u \end{bmatrix} = H^{(o)} \begin{bmatrix} u_1 \\ i_2 \\ e_o \\ i_o \end{bmatrix} = \begin{bmatrix} H_{11}^{(o)} & H_{12}^{(o)} \\ H_{21}^{(o)} & H_{22}^{(o)} \end{bmatrix} \begin{bmatrix} u_1 \\ i_2 \\ e_o \\ i_o \end{bmatrix}.$$

Dabei sind die vorkommenden Operatoren durch die entsprechenden Matrizen von Übertragungsfunktionen definiert (entweder direkt im Frequenzbereich oder im Zeitbereich durch einen Satz von gewöhnlichen linearen Differentialgleichungen mit konstanten Koeffizienten). Aus Gl. (6) folgt eine Gleichung, die das Verhalten des linearen Netzwerkanteils bezüglich der n-Tore, an denen die Nichtlinearitäten liegen, vollständig beschreibt:

$$\begin{bmatrix} i_1 \\ u_2 \end{bmatrix} = H_{11}^{(o)} \begin{bmatrix} u_1 \\ i_2 \end{bmatrix} + H_{12}^{(o)} \begin{bmatrix} e_o \\ i_o \end{bmatrix}. \quad (7)$$

Um eine Gleichung für das gesamte Netzwerk formulieren zu können, benötigen wir noch die Beziehungen des nichtlinearen Netzwerkanteil

$$\begin{bmatrix} i_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} -G_1 u_1 \\ -R_2 i_2 \end{bmatrix}$$

oder kurz

$$\begin{bmatrix} i_1 \\ u_2 \end{bmatrix} = -N \begin{bmatrix} u_1 \\ i_2 \end{bmatrix} \quad (8)$$

mit $N \in \Psi(\alpha, \beta)$.

Dabei nehmen wir der Einfachheit halber an, daß $\mathbf{a} = (\tilde{\alpha}, \tilde{\alpha}, \dots, \tilde{\alpha})^T$ mit $\tilde{\alpha} > 0$ und $\mathbf{\beta} = (\tilde{\beta}, \tilde{\beta}, \dots, \tilde{\beta})^T$ gilt, was wegen der vorausgesetzten Passivität von $\mathbf{N}$ durch geeignete Wahl von $\tilde{\alpha}$ und $\tilde{\beta}$ immer erreicht werden kann. Aus Gl. (7) und Gl. (8) folgt schließlich die Gleichung

$$-\mathbf{H}_{12}^{(o)} \begin{bmatrix} e_0 \\ i_0 \end{bmatrix} = \mathbf{N} \begin{bmatrix} u_1 \\ i_2 \end{bmatrix} + \mathbf{H}_{11}^{(o)} \begin{bmatrix} u_1 \\ i_2 \end{bmatrix},$$

die mit den Abkürzungen

$$\mathbf{f} = \begin{bmatrix} u_1 \\ i_2 \end{bmatrix}, \quad \mathbf{e} = -\mathbf{H}_{12}^{(o)} \begin{bmatrix} e_0 \\ i_0 \end{bmatrix} \quad \text{und} \quad \mathbf{H} = \mathbf{H}_{11}^{(o)}$$

eine Operatorgleichung unseres Standardtyps

$$\mathbf{e} = \mathbf{N}\mathbf{f} + \mathbf{H}\mathbf{f} \tag{9}$$

darstellt.

Wegen $\mathbf{T}_0 = \mathbf{I}$ wären die Voraussetzungen von Satz II.30 erfüllt, falls

$$\mathbf{e} \in \mathbf{L}_{2r}^n(-\infty, \infty)$$

und die positive, reelle und rationale Matrix $\mathbf{H}_{11}^{(o)}(p)$ imaginärreduziert wäre (= keine Pole auf der imaginären Achse). Es wäre dann nämlich $\mathbf{H} \in \mathbf{F}$, und es gelte

$$\mathbf{H}_{11}^{(o)H}(j\omega) + \mathbf{H}_{11}^{(o)}(j\omega) \geq 0 \quad \text{für alle } \omega \in \mathbf{R}.$$

Da dies aber im allgemeinen nicht der Fall ist, wollen wir durch eine geeignete Transformation von Gl. (9) diese Schwierigkeit umgehen. Bei dieser Transformation wird vom nichtlinearen Operator $\mathbf{N}$ ein "kleiner" linearer Operator $\epsilon \mathbf{I}$ (mit $\epsilon > 0$) abgezogen und dieser gleich-

zeitig dem linearen Operator hinzugefügt. Es entsteht somit die Gleichung

$$-\mathbf{H}_{12}^{(0)} \begin{bmatrix} e_0 \\ i_0 \end{bmatrix} = (\mathbf{N} - \epsilon \mathbf{I}) \begin{bmatrix} u_1 \\ i_2 \end{bmatrix} + (\mathbf{H}_{11}^{(0)} + \epsilon \mathbf{I}) \begin{bmatrix} u_1 \\ i_2 \end{bmatrix}$$

und, falls $(\mathbf{H}_{11}^{(0)} + \epsilon \mathbf{I})^{-1}$ existiert, gilt

$$-(\mathbf{H}_{11}^{(0)} + \epsilon \mathbf{I})^{-1} \mathbf{H}_{12}^{(0)} \begin{bmatrix} e_0 \\ i_0 \end{bmatrix} = (\mathbf{H}_{11}^{(0)} + \epsilon \mathbf{I})^{-1} (\mathbf{N} - \epsilon \mathbf{I}) \begin{bmatrix} u_1 \\ i_2 \end{bmatrix} + \begin{bmatrix} u_1 \\ i_2 \end{bmatrix}. \quad (10)$$

Dabei muß gewährleistet sein, daß

$$\mathbf{H}_{12}^{(0)} \begin{bmatrix} e_0 \\ i_0 \end{bmatrix}$$

im Definitionsbereich von $(\mathbf{H}_{11}^{(0)} + \epsilon \mathbf{I})^{-1}$ enthalten ist (daß dies der Fall ist, wird später gezeigt). Mit den Abkürzungen

$$\hat{\mathbf{H}} = (\mathbf{H}_{11}^{(0)} + \epsilon \mathbf{I})^{-1},$$

$$\hat{\mathbf{N}} = \mathbf{N} - \epsilon \mathbf{I},$$

$$\mathbf{f} = \begin{bmatrix} u_1 \\ i_2 \end{bmatrix}$$

und

$$\hat{\mathbf{e}} = -(\mathbf{H}_{11}^{(0)} + \epsilon \mathbf{I})^{-1} \mathbf{H}_{12}^{(0)} \begin{bmatrix} e_0 \\ i_0 \end{bmatrix}$$

geht Gl. (10) in unsere Standardgleichung

$$\hat{\mathbf{e}} = \mathbf{f} + \hat{\mathbf{H}} \hat{\mathbf{N}} \mathbf{f} \quad (11)$$

über.

Um zu beweisen, daß $(\mathbf{H}_{11}^{(0)} + \varepsilon \mathbf{I})^{-1}$ existiert, müssen wir uns noch über den Definitions- und Bildbereich von $(\mathbf{H}_{11}^{(0)} + \varepsilon \mathbf{I})$ unterhalten. Wie $\mathbf{H}_{11}^{(0)}$ ist auch $(\mathbf{H}_{11}^{(0)} + \varepsilon \mathbf{I})$ durch einen Satz von gewöhnlichen linearen Differentialgleichungen mit konstanten Koeffizienten gegeben, der aus der Hybridmatrix $(\mathbf{H}_{11}^{(0)}(p) + \varepsilon \mathbf{1}_n)$ abgeleitet werden kann. Diese Differentialgleichungen können (z.B. nach Transformation in den Zustandsraum mit zum Zeitpunkt $t = -\infty$ verschwindendem Zustandsvektor) für eine gewisse Menge Ω_1 von Erregungsfunktionen (definiert für alle $t \in \mathbf{R}$) gelöst werden, und wir nennen Ω_2 die Menge der Funktionen, die als Lösungen des Differentialgleichungssystems auftreten, d.h. $(\mathbf{H}_{11}^{(0)} + \varepsilon \mathbf{I}) : \Omega_1 \xrightarrow{\text{auf}} \Omega_2$.

Nach Newcomb[6] existiert die Hybridmatrix $(\mathbf{H}_{11}^{(0)}(p) + \varepsilon \mathbf{1}_n)^{-1}$ und damit ein zum Differentialgleichungssatz von $(\mathbf{H}_{11}^{(0)} + \varepsilon \mathbf{I})$ inverser Satz von Differentialgleichungen, der den Operator $(\mathbf{H}_{11}^{(0)} + \varepsilon \mathbf{I})^{-1}$ definiert, und es gilt $(\mathbf{H}_{11}^{(0)} + \varepsilon \mathbf{I})^{-1} : \Omega_2 \xrightarrow{\text{auf}} \Omega$.

Da die Matrix $\hat{\mathbf{H}}(p) = (\mathbf{H}_{11}^{(0)}(p) + \varepsilon \mathbf{1}_n)^{-1}$ ferner positiv, reell, rational und imaginärreduziert ist[6], bildet der Operator $\hat{\mathbf{H}}$ den Raum $\mathbf{L}_{2r}^n(-\infty, \infty)$ in sich ab, er ist beschränkt, und es gilt $\hat{\mathbf{H}}\varepsilon\mathbf{F}$ und $\hat{\mathbf{H}}^H(j\omega) + \hat{\mathbf{H}}(j\omega) \geq 0$ für alle $\omega \in \mathbf{R}$. Ferner ist $\hat{\mathbf{N}}\varepsilon\psi(\hat{\alpha}, \hat{\beta})$ mit $\hat{\alpha} = (\tilde{\alpha} - \varepsilon, \tilde{\alpha} - \varepsilon, \dots, \tilde{\alpha} - \varepsilon)^T$ und $\hat{\beta} = (\tilde{\beta} - \varepsilon, \tilde{\beta} - \varepsilon, \dots, \tilde{\beta} - \varepsilon)^T$.

Wenn wir jetzt noch annehmen, daß $\hat{\mathbf{N}}$ passiv ist (was mit der Wahl $0 < \varepsilon < \tilde{\alpha}$ immer erreicht werden kann), und uns vor Augen halten, daß $\hat{\mathbf{T}}_0 = \mathbf{I}$ ist, wäre Satz II.29 auf die Operatorgleich (11) anwendbar, falls

$$\hat{\mathbf{e}} = -(\mathbf{H}_{11}^{(0)} + \varepsilon \mathbf{I})^{-1} \mathbf{H}_{12}^{(0)} \begin{bmatrix} \mathbf{e}_0 \\ \mathbf{i}_0 \end{bmatrix} \varepsilon \mathbf{L}_{2r}^n(-\infty, \infty)$$

gelte. Um zu beweisen, daß dies der Fall ist, machen wir uns zunächst klar, was die Transformation von Gl. (9) in Gl. (11) bedeutet. Offenbar wird von jedem der nichtlinearen zeitvarianten Widerstände bzw. Leitwerte der Wert ϵ extrahiert und dem linearen Netzwerk zugefügt. (Das Gesamtnetzwerk bleibt damit unverändert.) Wir haben also die Situation von Abb. (4), wobei das erweiterte lineare Netzwerk durch die Hybridmatrix $H^{(1)}(p)$ beschrieben werden kann.

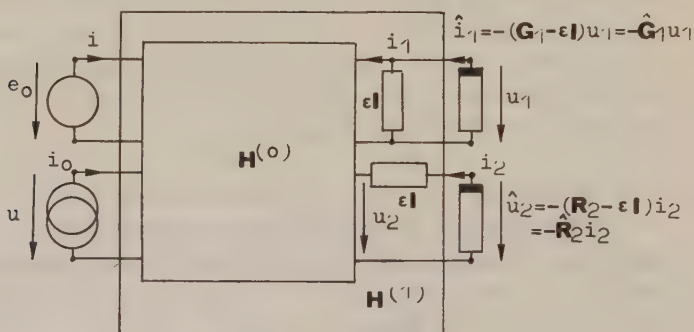


Abb. (4)

$$\begin{bmatrix} U_1(p) \\ I_2(p) \\ \hline I(p) \\ U(p) \end{bmatrix} = H^{(1)}(p) \begin{bmatrix} \hat{I}_1(p) \\ \hat{U}_2(p) \\ \hline E_o(p) \\ I_o(p) \end{bmatrix} = \begin{bmatrix} H_{11}^{(1)}(p) & H_{12}^{(1)}(p) \\ \hline H_{21}^{(1)}(p) & H_{22}^{(1)}(p) \end{bmatrix} \begin{bmatrix} \hat{I}_1(p) \\ \hat{U}_2(p) \\ \hline E_o(p) \\ I_o(p) \end{bmatrix}.$$

Dabei ist

$$H_{11}^{(1)}(p) = (H_{11}^{(0)}(p) + \epsilon 1_n)^{-1},$$

$$H_{12}^{(1)}(p) = -(H_{11}^{(0)}(p) + \epsilon 1_n)^{-1} H_{12}^{(0)}(p),$$

$$H_{21}^{(1)}(p) = H_{21}^{(0)}(H_{11}^{(0)}(p) + \epsilon I_n)^{-1} \quad \text{und}$$

$$H_{22}^{(1)}(p) = H_{22}^{(0)}(p) - H_{21}^{(0)}(p)(H_{11}^{(0)}(p) + \epsilon I_n)^{-1}H_{12}^{(0)}(p).$$

$H^{(1)}(p)$ ist eine positive, reelle und rationale Matrix, und da $(H_{11}^{(0)}(p) + \epsilon I_n)^{-1}$ imaginärreduziert ist, hat auch die Matrix $(H_{11}^{(0)}(p) + \epsilon I_n)^{-1}H_{12}^{(0)}(p)$ nur in der offenen linken p -Halbebene Pole [6]. Damit bildet der Operator $(H_{11}^{(0)} + \epsilon I)^{-1}H_{12}^{(0)}$ den Raum $L_{2r}^m(-\infty, \infty)$ in den Raum $L_{2r}^n(-\infty, \infty)$ ab, womit bewiesen ist, daß $\hat{e} \in L_{2r}^n(-\infty, \infty)$ gilt. Damit ist auch gezeigt, daß

$$H_{12}^{(0)} \begin{bmatrix} e_0 \\ i_0 \end{bmatrix}$$

im Definitionsbereich von $(H_{11}^{(0)} + \epsilon I)^{-1}$ liegt.

Jetzt sind alle Voraussetzungen und Bedingungen von Satz II.29 bezüglich Gl. (11) erfüllt, und folglich existiert ein eindeutiges $f \in L_{2r}^n(-\infty, \infty)$, das Gl. (11) befriedigt. Es bleibt noch zu zeigen, daß dieses f auch unsere ursprüngliche Gl. (9) bezüglich $L_{2r}^n(-\infty, \infty)$ eindeutig erfüllt.

Lemma 1: Es seien die Voraussetzungen dieses Abschnittes gegeben. Falls $f \in L_{2r}^n(-\infty, \infty)$ eine Lösung der Gleichung

$$e = Nf + Hf \tag{a}$$

dargestellt, ist f auch eine Lösung der Gleichung

$$(H + \epsilon I)^{-1}e = (H + \epsilon I)^{-1}(N - \epsilon I)f + f \tag{b}$$

und umgekehrt. (Es muß dabei nicht notwendig $e \in L_{2r}^n(-\infty, \infty)$ gelten.)

Beweis:

(1) (a) $\longrightarrow$ (b)

Es ist also $e = Nf + Hf$ mit $f \in L_{2r}^n(-\infty, \infty)$ gegeben.
Dann gilt $e = (H + \epsilon I)f + (N - \epsilon I)f$. Der Operator $(H + \epsilon I)^{-1}$
ist auf $(H + \epsilon I)f$ und $(N - \epsilon I)f$ und damit auch auf e
anwendbar. Also ist

$$(H + \epsilon I)^{-1} e = f + (H + \epsilon I)^{-1} (N - \epsilon I) f ,$$

und es gilt (b).

(2) (b) $\longrightarrow$ (a)

Es ist also $(H + \epsilon I)^{-1} e = f + (H + \epsilon I)^{-1} (N - \epsilon I)f$ mit
 $f \in L_{2r}^n(-\infty, \infty)$. Nach Anwendung des Operators $(H + \epsilon I)$
folgt (a).

q.e.d.

Wir haben also die Eindeutigkeit von $f \in L_{2r}^n(-\infty, \infty)$ be-
züglich Gl. (9) gezeigt, und damit gilt schließlich

Satz 2: Gegeben sei ein elektrisches Netzwerk gemäß den
Annahmen dieses Abschnitts (siehe Abb. (3)).
Dann existiert für alle Erregungen

$$\begin{bmatrix} e \\ 0 \\ i \\ 0 \end{bmatrix} \in L_{2r}^m(-\infty, \infty)$$

eine eindeutige Vektorfunktion

$$\begin{bmatrix} u_1 \\ i_2 \end{bmatrix} \in L_{2r}^n(-\infty, \infty)$$

der Ströme bzw. Spannungen durch bzw. an den
Nichtlinearitäten. Diese Vektorfunktion kann
mittels eines Iterationsverfahrens berechnet
werden.

3.2 Elektrische Netzwerke mit nichtlinearen zeitvarianten gesteuerten Quellen

Gegeben sei ein elektrisches Netzwerk, das lineare zeitinvariante (aber nicht notwendig konzentrierte, passive und reziproke) Elemente und ferner n nichtlineare zeitvariante gesteuerte Quellen enthält. Alle Energiespeicher seien zum Zeitpunkt $t = -\infty$ ungeladen. Erregt werde dieses Netzwerk durch Strom- und Spannungsquellen mit quadratisch integrierbarem zeitlichen Verlauf. Durch diese Quellen können auch die Anfangsbedingungen der Energiespeicher berücksichtigt werden[23]. Wir haben also die in Abb. (4) gezeigte Situation, wobei die eingezeichneten Größen folgende Bedeutung haben:

$e_0 \in \mathbb{L}_{2r}^r(-\infty, \infty)$ Vektorfunktion der r Spannungsquellen,

$i_0 \in \mathbb{L}_{2r}^s(-\infty, \infty)$ Vektorfunktion der s Stromquellen,

i, u Vektorfunktion der a gesuchten Ströme und b gesuchten Spannungen.

Die Vektorfunktionen der elektrischen Größen an den gesteuerten Quellen sind:

u_1, u_3 u spannungsgesteuerte Spannungsquellen,

i_1, u_4 v stromgesteuerte Spannungsquellen,

u_2, i_3 w spannungsgesteuerte Stromquellen,

i_2, i_4 c stromgesteuerte Stromquellen.

Es sei $n = u+v+w+c$, $d = a+b$ und $m = r+s$.

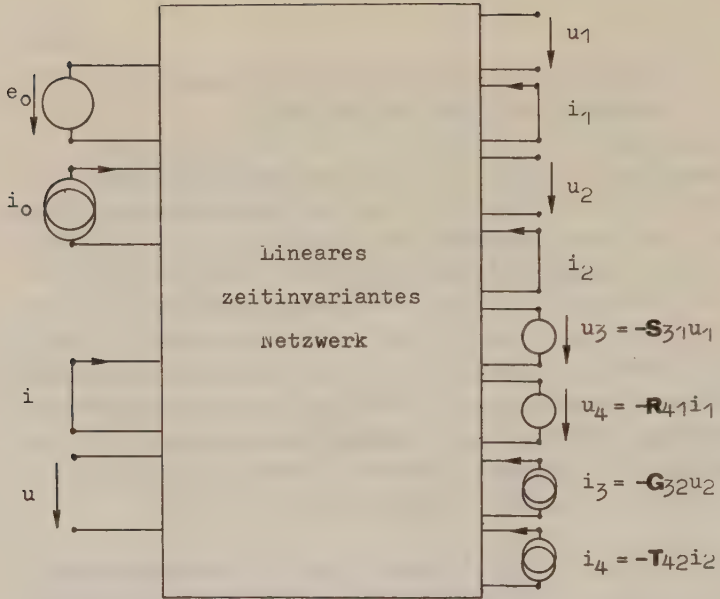


Abb. (4)

Wir nehmen an, daß im linearen Netzwerkanteil das Thevenin-Norton-Theorem in der Form

$$\begin{bmatrix} u_1 \\ i_1 \\ u_2 \\ i_2 \end{bmatrix} = \mathbf{H} \begin{bmatrix} u_3 \\ u_4 \\ i_3 \\ i_4 \end{bmatrix} + \mathbf{U} \begin{bmatrix} e_o \\ i_o \end{bmatrix} \quad (12)$$

für alle

$$\begin{bmatrix} u_3 \\ u_4 \\ i_3 \\ i_4 \end{bmatrix} \in \mathbf{L}_{2r}^n(-\infty, \infty)$$

und

$$\begin{bmatrix} e_o \\ i_o \end{bmatrix} \in \mathbf{L}_{2r}^m(-\infty, \infty)$$

gilt. (Dabei sind neben e_0 und i_0 auch u_3 , u_4 , i_3 und i_4 als linear unabhängige Quellen aufzufassen.) Die linearen Operatoren $\mathbf{H}$ und $\mathbf{U}$ seien Faltungsoperatoren, d.h. es ist $\mathbf{H}, \mathbf{U} \in \mathbf{F}$.

Die Eigenschaften der nichtlinearen zeitvarianten Quellen seien durch

$$u_{3i}(t) = -s_i(u_{1i}(t), t) \quad \text{für } i = 1, \dots, u,$$

$$u_{4i}(t) = -r_i(i_{1i}(t), t) \quad \text{für } i = 1, \dots, v,$$

$$i_{3i}(t) = -g_i(u_{2i}(t), t) \quad \text{für } i = 1, \dots, w \text{ und}$$

$$i_{4i}(t) = -t_i(i_{2i}(t), t) \quad \text{für } i = 1, \dots, c$$

oder kurz

$$u_3 = -\mathbf{S}_{31}u_1,$$

$$u_4 = -\mathbf{R}_{41}i_1,$$

$$i_3 = -\mathbf{G}_{32}u_2 \text{ und}$$

$$i_4 = -\mathbf{T}_{42}i_2$$

definiert, wobei mit

$$\begin{bmatrix} u_3 \\ u_4 \\ i_3 \\ i_4 \end{bmatrix} = - \begin{bmatrix} \mathbf{S}_{31}u_1 \\ \mathbf{R}_{41}i_1 \\ \mathbf{G}_{32}u_2 \\ \mathbf{T}_{42}i_2 \end{bmatrix} = -\mathbf{N} \begin{bmatrix} u_1 \\ i_1 \\ u_2 \\ i_2 \end{bmatrix} \quad (13)$$

für den nichtlinearen Operator $\mathbf{N}$ gemäß Abschnitt II.2.3 $\mathbf{N} \in \mathbf{N}(\alpha, \beta)$ gelte. Aus Gl. (12) und Gl. (13) folgt schließlich die Gleichung

$$\mathbf{U} \begin{bmatrix} e_o \\ i_o \end{bmatrix} = \begin{bmatrix} u_1 \\ i_1 \\ u_2 \\ i_2 \end{bmatrix} + \mathbf{HN} \begin{bmatrix} u_1 \\ i_1 \\ u_2 \\ i_2 \end{bmatrix},$$

die mit den Abkürzungen

$$\mathbf{e} = \mathbf{U} \begin{bmatrix} e_c \\ i_o \end{bmatrix}$$

und

$$\mathbf{f} = \begin{bmatrix} u_1 \\ i_1 \\ u_2 \\ i_2 \end{bmatrix}$$

in eine Gleichung unseres Standardtyps

$$\mathbf{e} = \mathbf{f} + \mathbf{HNf} \quad (14)$$

übergeht. Wegen $\mathbf{U} \in \mathbf{F}$ gilt $\mathbf{e} \in \mathbf{L}_{2r}^n(-\infty, \infty)$, und damit kann Satz II.27 auf Gl. (14) angewandt werden, falls die Bedingungen

- (i) $\det (1_{n+H(j\omega)C_o}) \neq 0$ für alle $\omega \in \mathbf{R}$,
- (ii) $\det (1_{n+h_o C_o}) \neq 0$ und
- (iii) $r_o = \sup_{\omega \in \mathbf{R}} \wedge \{ (1_{n+H(j\omega)C_o})^{-1} H(j\omega) T_o^{-1} \} \eta(C_o, T_o) < 1$

erfüllt sind. Gemäß Satz II.27 existiert dann ein eindeutiges $\mathbf{f} \in \mathbf{L}_{2r}^n(-\infty, \infty)$, das Gl. (14) befriedeigt. Diese Lösung kann anhand eines Iterationsverfahrens berechnet werden.

Die gesuchten Ströme i und Spannungen u des Netzwerks lassen sich schließlich mit bekannten Methoden aus der linearen Netzwerktheorie berechnen, und es gilt auf Grund des Kompensationstheorems

$$r = \begin{bmatrix} u \\ i \end{bmatrix} = \mathbf{V} \begin{bmatrix} u_3 \\ u_4 \\ i_3 \\ i_4 \end{bmatrix} + \mathbf{S} \begin{bmatrix} e_o \\ i_o \end{bmatrix} .$$

Falls $\mathbf{V}, \mathbf{S} \in \mathbf{F}$ ist, gilt außerdem $r \in \mathbf{L}_{2r}^d(-\infty, \infty)$.

Das gesamte Netzwerkverhalten kann also bezüglich der uns interessierenden Größen durch das Systemblockschaltbild (Abb. (5)) dargestellt werden.

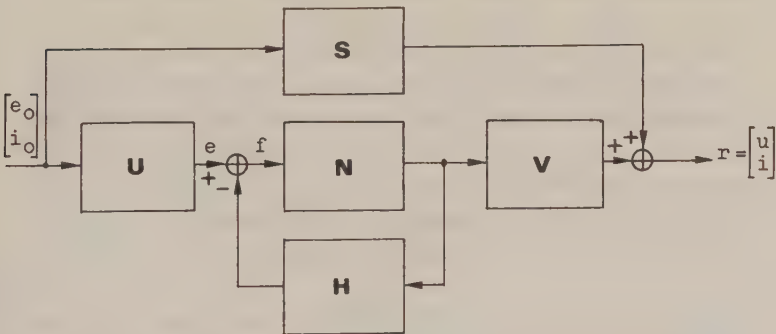


Abb. (5)

3.3 Elektrische Netzwerke mit nichtlinearen zeitvarianten Widerständen und gesteuerten Quellen

Häufig tritt der Fall auf, daß in einem elektrischen Netzwerk neben linearen zeitinvarianten Elementen sowohl Widerstände (bzw. Leitwerte) als auch gesteuerte Quellen enthalten sind, die nichtlinear und zeitvariant sind. Die Behandlung dieses Falles entspricht einer Kombination der Abschnitte III.3.1 und III.3.2 und führt bei analoger Vorgehensweise ebenfalls auf Gleichungen unseres Standardtyps.

3.4 Gleichstromanalyse elektrischer Netzwerke

Gegeben sei ein elektrisches Netzwerk, das neben linearen zeitinvarianten Elementen noch nichtlineare zeitinvariante Widerstände (bzw. Leitwerte) und gesteuerte Quellen enthält. Bei der Gleichstromanalyse sind die Erregungen des Netzwerks Gleichströme bzw. -spannungen, und man nimmt den eingeschwungenen Zustand an. (Dazu werden z.B. alle Induktivitäten durch Kurzschlüsse und alle Kapazitäten durch Leerläufe ersetzt.) Wie in den vorhergehenden Abschnitten erhält man bei der Analyse Gleichungen der Art

$$e = f + H N f \quad (15)$$

und

$$e = N f + H f, \quad (16)$$

wobei e den Vektor der Ersatzspannungs- bzw. Ersatzstromquellen (Gleichgrößen) und f den Vektor der Gleichspannungen bzw. -ströme an bzw. in den nichtlinearen Elementen darstellt.

Um die Theorie von Kapitel II. auf diese Gleichungen anwenden zu können, müssen wir noch gewisse Voraussetzungen treffen. Als Banachraum $\mathbf{R}^n$ von Kapitel II. wählen wir den Raum $\mathbf{R}^n$ mit irgendeiner Vektornorm (z.B. die euklidische Vektornorm). Der Operator

$\mathbf{H}: \mathbf{R}^n \rightarrow \mathbf{R}^n$ ist dann durch die reelle $n \times n$ -Matrix $\mathbf{H}$ definiert. Es seien $\psi_1(u), \psi_2(u), \dots, \psi_n(u)$ reellwertige Funktionen der reellen Variablen u definiert auf $\mathbf{R}$, und es existieren Vektoren $\alpha, \beta \in \mathbf{R}^n$ derart, daß mit $\alpha_i < \beta_i$ für $i = 1, \dots, n$ die Ungleichungen

$$\alpha_i \leq \frac{\psi_i(u) - \psi_i(v)}{u - v} \leq \beta_i$$

für $i = 1, \dots, n$ und alle $u, v \in \mathbf{R}$ mit $u \neq v$ gelten. Der nichtlineare Operator $\mathbf{H}: \mathbf{R}^n \rightarrow \mathbf{R}^n$ wird durch

$$(\psi h)_i = \psi_i(h_i) \quad (17)$$

für $i = 1, \dots, n$ mit beliebigen $h \in \mathbf{R}^n$ definiert, und wir schreiben $\mathbf{N}h = \psi(h)$, wobei $\psi(h)$ eine Vektorfunktion gemäß Gl. (17) ist. Gl. (15) bzw. Gl. (16) lauten jetzt in konventioneller Schreibweise

$$e = f + \mathbf{H}\psi(f) \quad (18)$$

bzw.

$$e = \psi(f) + \mathbf{H}f. \quad (19)$$

Es läßt sich leicht zeigen, daß mit den hier getroffenen Vereinbarungen alle Voraussetzungen der Abschnitte II.2.1 und II.2.2 erfüllt sind, und somit können die Sätze II.2 bzw. II.3 auf die Gl. (18) bzw. (19) angewandt werden. Natürlich werden wir auch hier die im Sinne von Abschnitt II.2.4.3 optimalen Transformationsoperatoren $\mathbf{C}_0$ und $\mathbf{T}_0$ bzw. die diesen Operatoren zugeordneten Matrizen $\mathbf{C}_0$ und $\mathbf{T}_0$ verwenden (siehe Abschnitt III.2). Damit kommen wir zu den folgenden Sätzen, deren Beweis trivial ist:

Satz 3: Es seien die Voraussetzungen dieses Abschnitts erfüllt, und es gelte

$$(i) \det(1_n + HC_0) \neq 0 \text{ und}$$

$$(ii) r_0 = \|(1_n + HC_0)^{-1} H T_0^{-1}\| \eta(C_0, T_0) < 1.$$

Dann gibt es für jedes $e \in \mathbf{R}^n$ ein eindeutiges $f \in \mathbf{R}^n$, so daß $e = f + H\psi(f)$ gilt. Ferner ist

$$f = \lim_{\mu \rightarrow \infty} f_\mu,$$

wobei

$$f_{\mu+1} = -(1_n + HC_0)^{-1} H(\psi(f_\mu) - C_0 f_\mu) + (1_n + HC_0)^{-1} e$$

gilt und f_0 ein beliebiges Element aus $\mathbf{R}^n$ ist.

Weiter gilt für die μ -te Approximation f_μ von f

$$\|f_\mu - f\| \leq \frac{r_0^\mu}{1 - r_0} \|f_1 - f_0\|.$$

Satz 4: Es seien die Voraussetzungen dieses Abschnitts erfüllt, und es gelte

$$(i) \det(H + C_0) \neq 0 \text{ und}$$

$$(ii) q_0 = \|(H + C_0)^{-1} T_0^{-1}\| \eta(C_0, T_0) < 1.$$

Dann gibt es für jedes $e \in \mathbf{R}^n$ ein eindeutiges $f \in \mathbf{R}^n$, so daß $e = \psi(f) + Hf$ gilt. Ferner ist

$$f = \lim_{\mu \rightarrow \infty} f_\mu,$$

wobei

$$f_{\mu+1} = (H + C_0)^{-1} (\psi(f_\mu) - C_0 f_\mu) + (H + C_0)^{-1} e$$

gilt und f_0 ein beliebiges Element aus $\mathbf{R}^n$ ist.

Weiter gilt für die μ -te Approximation f_μ von f

$$\|f_\mu - f\| \leq \frac{q_0^\mu}{1 - q_0} \|f_1 - f_0\|.$$

Anhand dieser Sätze sind wir nun in der Lage, die Gl. (18) und Gl. (19), die sich bei der Gleichstromanalyse von elektrischen Netzwerken ergeben, hinsichtlich ihrer eindeutigen Lösbarkeit zu untersuchen. Ferner können wir die gesuchten Lösungen mittels eines Iterationsverfahrens berechnen.

Bemerkung:

Die in den Sätzen 3 und 4 vorkommenden Matrixnormen sind beliebig, da ja auch die Norm des Raumes $\mathbb{R}^n$ willkürlich wählbar war. (Die Matrixnorm muß aber mit der Vektornorm verträglich sein.)

IV. NUMERISCHE ERGEBNISSE

1. Einleitung

An die mehr mathematisch orientierte Untersuchung systemtheoretischer Probleme in Kapitel II. schließt sich jetzt die Frage nach der günstigen Ermittlung numerischer Resultate an.

In diesem Kapitel werden wir die theoretischen Ergebnisse des Kapitels II. auf konkrete Analyseprobleme aus der Systemtheorie (wie z.B. die Analyse elektrischer Netzwerke gemäß Kapitel III.) praktisch anwenden. Bei unseren Berechnungen beschränken wir uns auf Zeitfunktionen des Raumes $L_{2r}^n(-\infty, \infty)$, da mit Funktionen dieser Klasse sowohl wichtige praktische Fragestellungen als auch interessante theoretische Aussagen formulierbar sind. Es wird ein numerisches Verfahren vorgeschlagen, das es gestattet, die Lösung der das System beschreibenden nichtlinearen Integralgleichung iterativ zu berechnen, falls ein Kriterium für die Existenz und Eindeutigkeit der Lösung erfüllt ist. Das Analyseproblem besteht im folgenden nicht nur aus einer Berechnung der Lösung, sondern auch in der Absicherung ihrer Eindeutigkeit. Modelle von physikalischen Anordnungen sind nur bei eindeutiger Lösbarkeit der Modellgleichungen sinnvoll, da ja der Ablauf eines reproduzierbaren physikalischen Experiments auch eindeutig ist und eindeutige Meßergebnisse liefert (abgesehen von Meßfehlern).

Die Leistungsfähigkeit des Verfahrens wird anhand verschiedener Analysebeispiele demonstriert. Ferner wird das Verfahren für den Fall, daß das System aus konzentrierten Elementen besteht, mit einem bekannten Integrationsverfahren verglichen.

2. Ein Analyseverfahren

2.1 Voraussetzungen und Problemstellung

Das zu analysierende System sei gemäß Abb. (1) aufgebaut. Wie in Kapitel III. gezeigt wurde, lassen sich z.B. elektrische Netzwerke mit nichtlinearen und zeitvarianten Widerständen und gesteuerten Quellen durch Systeme dieser Art beschreiben. Dabei sei $e_0 \in L_{2r}^q(-\infty, \infty)$ die Systemerregung und r die Systemreaktion.

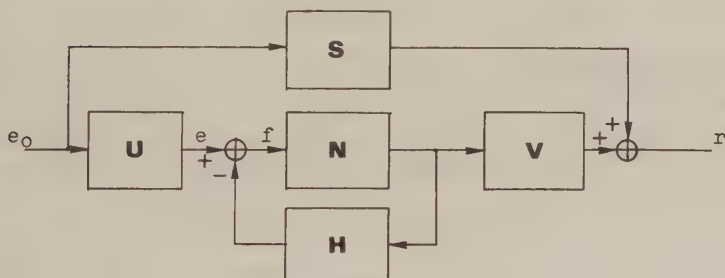


Abb.(1)

Wir nehmen an, daß die linearen Operatoren **U**, **S**, **H** und **V** vom Faltungstyp sind und daß

$$\mathbf{N} \in \Psi(\alpha, \beta)$$

gilt (siehe I.4.4 und II.2.3). Ferner seien die Operatoren **U**, **S**, **H**, **V** und $(\mathbf{I} + \mathbf{H}\mathbf{C}_0)^{-1}$ im Sinne von Abschnitt I.2.6.3 kausal (auf Grund der Definition sind die Operatoren **N** und **C**₀ von vornherein kausal), was bei anwendungsbezogenen Aufgaben keine Einschränkung bedeutet, da diesen Operatoren dann konkrete Systeme entsprechen. Insbesondere stellt der Operator $(\mathbf{I} + \mathbf{H}\mathbf{C}_0)^{-1}$ das durch $\mathbf{N} = \mathbf{C}_0$ linearisierte Ersatzsystem dar.

Die Faltungsoperatoren **U**, **S**, **H** und **V** seien gemäß Abschnitt I.4.4 im Frequenzbereich durch die Fourier-Transformierten der Impulsantworten $U(j\omega)$, $S(j\omega)$, $H(j\omega)$ und $V(j\omega)$ definiert.

Auf Grund der Eigenschaften von Faltungsoperatoren gilt

$$e \in L_{2r}^n(-\infty, \infty),$$

und damit reduziert sich das Analyseproblem im wesentlichen auf die Bestimmung von f . Wegen Satz II.27 existiert eine eindeutige Lösung $f \in L_{2r}^n(-\infty, \infty)$, falls die Bedingungen

- (i) $\det (1_n + H(j\omega)C_0) \neq 0$ für alle $\omega \in \mathbf{R}$;
- (ii) $\det (1_n + h_0 C_0) \neq 0$ und
- (iii) $r_0 < 1$

erfüllt sind. Die Lösung f kann dann anhand der Iterationsvorschrift

$$f = \lim_{\mu \rightarrow \infty} f_\mu$$

mit

$$f_{\mu+1} = -(I + HC_0)^{-1}H(N - C_0)f_\mu + (I + HC_0)^{-1}e$$

und beliebigem $f_0 \in L_{2r}^n(-\infty, \infty)$ berechnet werden, und es gilt

$$\|f - f_\mu\| \leq \frac{r_0^\mu}{1 - r_0} \|f_1 - f_0\|.$$

Diese Iterationsvorschrift wollen wir in einer leicht abgeänderten Form zur numerischen Berechnung der Lösung f verwenden, und wir setzen dabei grundsätzlich $f_0 = \theta$. Nach der Berechnung von f , die natürlich wie bei allen numerischen Verfahren nur näherungsweise erfolgen kann, sind wir in der Lage, die Systemreaktion r gemäß

$$r = Se_0 + Vnf$$

anzugeben, und es gilt

$$r \in L_{2r}^m(-\infty, \infty).$$

Bemerkung 1:

Nach dem Beweis von Satz II.23 ist der Operator $(I + HC_0)^{-1}$ im Frequenzbereich durch die Matrixfunktion

$$(1_n + H(j\omega)C_0)^{-1}$$

definiert.

Bemerkung 2:

Die Kausalität von Faltungsoperatoren ist dann gewährleistet, wenn die Impulsantwort für negative Zeiten verschwindet. Es kann gezeigt werden [16], daß der Operator $(I + HC_0)^{-1}$ kausal ist, falls für alle $\sigma, \omega \in \mathbf{R}$ mit $\sigma \geq 0$

$$(i) \quad \det(1_n + H(\sigma + j\omega)C_0) \neq 0 \quad \text{und}$$

$$(ii) \quad \det(1_n + h_0 C_0) \neq 0$$

gilt. Im rationalen Fall bedeutet dies, daß die Matrix $(1_n + H(p)C_0)^{-1}$ keine Pole in der abgeschlossenen rechten p -Halbebene (einschließlich $p = \infty$) besitzt.

Bemerkung 3:

Die Kausalitätsforderung an den Operator $(I + HC_0)^{-1}$ bedeutet praktisch, daß das linearisierte System stabil sein muß [16].

2.2 Iteration in Bezug auf einen endlichen Zeitbereich

In vielen praktisch interessanten Fällen nimmt man an, daß die Erregungsfunktion e_0 des Systems kausal ist, d.h. daß

$$e_0(t) = 0$$

für alle $t < 0$ gilt. Man wird auf Grund physikalischer Überlegungen erwarten, daß die anderen Systemfunktionen e , f und r ebenfalls kausal sind. Wie der folgende Satz zeigt, ist das wegen der Kausalität der Operatoren $\mathbf{U}$, $\mathbf{S}$, $\mathbf{H}$, $\mathbf{V}$, $\mathbf{C}_0$, $\mathbf{N}$ und $(\mathbf{I} + \mathbf{H}\mathbf{C}_0)^{-1}$ auch tatsächlich der Fall.

Satz 1: Es seien die Voraussetzungen und die Bedingungen (i), (ii) und (iii) von Abschnitt IV.2.1 erfüllt, und es gelte $e_0(t) = 0$ für alle $t < t_0$. Dann verschwinden auch die Systemfunktionen e , f und r für fast alle $t < t_0$.

Beweis:

Wegen der Kausalität von $\mathbf{U}$ gilt

$$\mathbf{P}_{t_0} e = \mathbf{P}_{t_0} \mathbf{U} e_0 = \mathbf{P}_{t_0} \mathbf{U} \mathbf{P}_{t_0} e_0 ,$$

und damit folgt aus $\mathbf{P}_{t_0} e_0 = 0$ sofort

$$\mathbf{P}_{t_0} e = 0 \tag{1}$$

bzw. $e(t) = 0$ für fast alle $t < t_0$. Es ist

$$f = -(\mathbf{I} + \mathbf{H}\mathbf{C}_0)^{-1} \mathbf{H}(\mathbf{N} - \mathbf{C}_0) f + (\mathbf{I} + \mathbf{H}\mathbf{C}_0)^{-1} e ,$$

und wegen der Kausalität aller Operatoren gilt

$$\mathbf{P}_{t_0} f = -\mathbf{P}_{t_0} (\mathbf{I} + \mathbf{H}\mathbf{C}_0)^{-1} \mathbf{H}(\mathbf{N} - \mathbf{C}_0) \mathbf{P}_{t_0} f + \mathbf{P}_{t_0} (\mathbf{I} + \mathbf{H}\mathbf{C}_0)^{-1} \mathbf{P}_{t_0} e .$$

Mit Gl. (1) ist

$$\mathbf{P}_{t_0} f = -\mathbf{P}_{t_0} (\mathbf{I} + \mathbf{H} \mathbf{C}_0)^{-1} \mathbf{H} (\mathbf{N} - \mathbf{C}_0) \mathbf{P}_{t_0} f.$$

Wegen

$$\begin{aligned} g_0(\mathbf{P}_{t_0} (\mathbf{I} + \mathbf{H} \mathbf{C}_0)^{-1} \mathbf{H} (\mathbf{N} - \mathbf{C}_0)) &\leq g_0((\mathbf{I} + \mathbf{H} \mathbf{C}_0)^{-1} \mathbf{H} (\mathbf{N} - \mathbf{C}_0)) \\ &\leq r(\mathbf{C}_0, \mathbf{T}_0) = r_0 < 1 \end{aligned}$$

folgt

$$\|\mathbf{P}_{t_0} f\| \leq r_0 \|\mathbf{P}_{t_0} f\|$$

und damit

$$\|\mathbf{P}_{t_0} f\| = 0 \text{ bzw. } \mathbf{P}_{t_0} f = \theta.$$

Damit ist der Nachweis erbracht, daß $f(t)$ für fast alle $t < t_0$ verschwindet. Es bleibt noch zu zeigen, daß dies auch für $r(t)$ gilt:

$$r = \mathbf{S} e_0 + \mathbf{V} \mathbf{N} f$$

$$\mathbf{P}_{t_0} r = \mathbf{P}_{t_0} \mathbf{S} \mathbf{P}_{t_0} e_0 + \mathbf{P}_{t_0} \mathbf{V} \mathbf{N} \mathbf{P}_{t_0} f = \theta.$$

q.e.d.

Der folgende Satz legitimiert eine Beschränkung des Zeitraums, in dem die gesuchten Funktionen berechnet werden, auf das Intervall $[0, T]$, wobei er auch die Anwendung eines modifizierten Iterationsverfahrens erlaubt, das für den praktischen Einsatz besser geeignet ist als das Iterationsverfahren von Abschnitt IV.2.1. Dabei können alle Funktionen außerhalb des Intervalls $[0, T]$ gleich Null gesetzt werden, und diese Funktionen (insbesondere die sich aus dem Iterationsprozeß ergebende Funktion) stellen dann im Zeitraum $[0, T]$ die gesuchten Funktionen e, f und r dar. Ferner ist die Wahl des Endzeitpunktes $T > 0$ willkürlich und kann z.B. nach numerischen Gesichtspunkten erfolgen.

Satz 2: Es seien die Voraussetzungen und Bedingungen (i), (ii) und (iii) von Abschnitt IV.2.1 erfüllt, und es gelte $e_0(t) = 0$ für alle $t < 0$. Schließlich seien die Funktionen $\hat{e}$, $\hat{e}$, $\hat{f}$ und $\hat{r}$ folgendermaßen gegeben (mit $T > 0$):

$$(a) \quad \hat{e}_0 = P_T e_0,$$

$$(b) \quad \hat{e} = P_T U \hat{e}_0 = P_T U P_T e_0,$$

$$(c) \quad \hat{f} = \lim_{\mu \rightarrow \infty} \hat{f}_\mu \text{ mit} \\ \hat{f}_{\mu+1} = -P_T(I + HC_0)^{-1}H(N-C_0)\hat{f}_\mu + P_T(I + HC_0)^{-1}\hat{e} \\ \text{für } \mu = 0, 1, 2, \dots \text{ und } \hat{f}_0 = \theta. \text{ (Diese Iterations-} \\ \text{vorschrift konvergiert.)}$$

$$(d) \quad \hat{r} = P_T S \hat{e}_0 + P_T V N \hat{f}.$$

Dann stimmen die Funktionswerte der Funktionen e_0 , e , f und r jeweils mit den Funktionswerten der Funktionen $\hat{e}_0$, $\hat{e}$, $\hat{f}$ und $\hat{r}$ für fast alle $t \in (-\infty, T)$ überein. Die Funktionswerte all dieser Funktionen verschwinden für fast alle negativen Zeiten, während die Funktionswerte von $\hat{e}_0$, $\hat{e}$, $\hat{f}$ und $\hat{r}$ auch für fast alle $t > T$ verschwinden.

Beweis:

Zunächst ist zu zeigen, daß die Iterationsvorschrift zur Berechnung von $\hat{f}$ konvergiert. Wegen

$$\begin{aligned} g_0(P_T(I + HC_0)^{-1}H(N-C_0)) &\leq g_0((I + HC_0)^{-1}H(N-C_0)) \\ &\leq r(C_0, T_0) = r_0 < 1 \end{aligned}$$

ist der durch

$$\Phi g = -P_T(I + HC_0)^{-1}H(N-C_0)g + P_T(I + HC_0)^{-1}\hat{e}$$

definierte Operator

$$\Phi : L_{2r}^n(-\infty, \infty) \longrightarrow L_{2r}^n(-\infty, \infty)$$

eine Kontraktion und deshalb existiert ein eindeutiges $\hat{f} \in L_{2r}^n(-\infty, \infty)$, für das

$$\hat{f} = \Phi \hat{f}$$

gilt, und das gemäß (c) ermittelt werden kann. Auf Grund der Definition ist klar, daß die Funktionswerte der Funktionen $\hat{e}_0$, $\hat{e}$, $\hat{f}$ und $\hat{r}$ für alle $t > T$ verschwinden. Wir zeigen jetzt, daß die Funktionswerte von e_0 , e , f und r jeweils mit den Funktionswerten von $\hat{e}_0$, $\hat{e}$, $\hat{f}$ und $\hat{r}$ für fast alle $t \leq T$ übereinstimmen:

Auf Grund der Definition ist

$$\hat{e}_0 = P_T e_0,$$

und wegen der Kausalität von U gilt

$$\hat{e} = P_T e.$$

Aus

$$f = -(I + HC_0)^{-1} H(N - C_0)f + (I + HC_0)^{-1} e$$

folgt unter Verwendung der Kausalität aller Operatoren

$$P_T f = -P_T (I + HC_0)^{-1} H(N - C_0) P_T f + P_T (I + HC_0)^{-1} P_T e$$

oder kurz

$$P_T f = \Phi P_T f.$$

Wegen der Eindeutigkeit des Fixpunktes von Φ gilt schließlich

$$\hat{f} = P_T f.$$

Die Gleichung

$$\hat{r} = P_T r$$

folgt mit

$$r = S e_0 + V N f$$

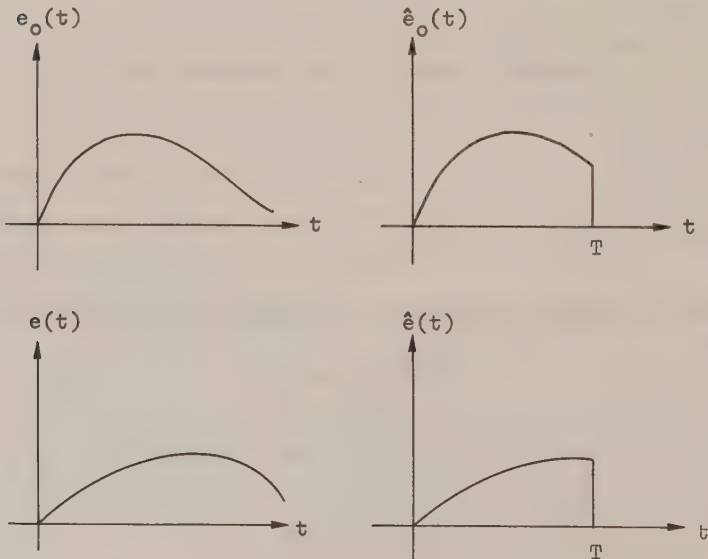
aus

$$\begin{aligned}
 \mathbf{P}_T \mathbf{r} &= \mathbf{P}_T \mathbf{S} \mathbf{e}_0 + \mathbf{P}_T \mathbf{V} \mathbf{N} \mathbf{f} \\
 &= \mathbf{P}_T \mathbf{S} \mathbf{P}_T \mathbf{e}_0 + \mathbf{P}_T \mathbf{V} \mathbf{N} \mathbf{P}_T \mathbf{f} \\
 &= \mathbf{P}_T \mathbf{S} \hat{\mathbf{e}}_0 + \mathbf{P}_T \mathbf{V} \mathbf{N} \hat{\mathbf{f}} \\
 &= \hat{\mathbf{r}}.
 \end{aligned}$$

Aus Satz 1 folgt mit $t_0 = 0$, daß die Funktionswerte von e , f und r für fast alle $t < 0$ verschwinden, und wegen $T > 0$ verschwinden damit die Funktionswerte aller Funktionen für fast alle $t < 0$.

q.e.d.

Diese Aussage von Satz 2 soll anhand eines Beispiels gemäß Abb. (2) verdeutlicht werden.



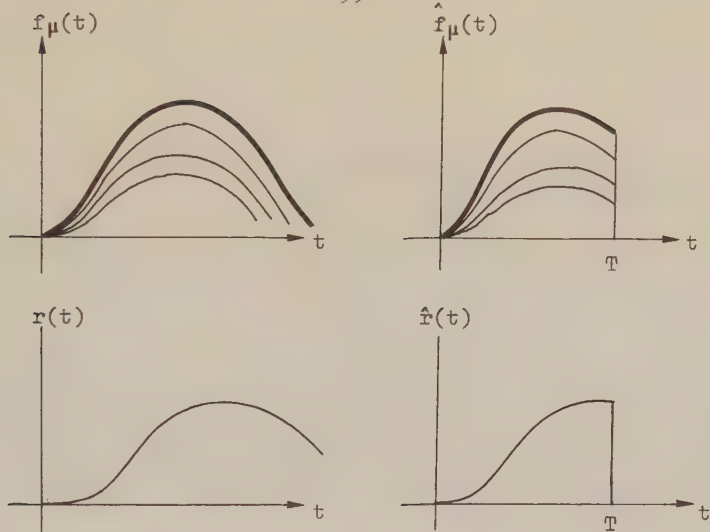


Abb.(2)

Da die Lösungsfunktion f quadratisch integrierbar ist, weist sie für $t \rightarrow \infty$ ein "abklingendes" Verhalten auf, und deshalb interessiert bei ihrer praktischen Berechnung häufig nur der Verlauf in einem endlichen Zeitintervall $[0, T]$. In unserem numerischen Verfahren werden daher sämtliche Berechnungen gemäß den Konstruktionsvorschriften (a), (b), (c) und (d) von Satz 2 durchgeführt. Ein weiterer Grund für die Berechnung der Funktionen $\hat{e}$, $\hat{f}$ und $\hat{r}$ anstelle von e , f und r liegt in der Tatsache, daß die Darstellung von Funktionen am Digitalrechner durch äquidistante Abtastwerte nur für endliche Zeitbereiche möglich ist (auf Grund des endlichen Speichers). Ferner besteht für diese "abgebrochenen" Funktionen die Möglichkeit der numerischen Fourier-Transformation.

2.3 Realisierung des Verfahrens unter Verwendung eines Digitalrechners

Es seien die Voraussetzungen von Abschnitt IV.2.1 gegeben, und es gelte für die Systemerregung

$$e_0(t) = 0$$

für alle $t < 0$. Der für die Analyse interessante Zeitbereich sei das Intervall $[0, T]$, und wir nehmen an, daß in diesem Intervall $e_0(t)$ eine bis auf endlich viele Sprungstellen glatte Funktion ist. (Das ist für den Anwender eine unwesentliche Einschränkung.)

In diesem Abschnitt werden wir zeigen, wie die Konstruktionsvorschriften (a), (b), (c) und (d) für die gesuchten "abgebrochenen" Funktionen $\hat{e}$, $\hat{f}$ und $\hat{r}$ mit Hilfe eines Digitalrechners realisiert werden können. Es liegt in der Natur numerischer Verfahren, daß die gesuchten Größen nur mit einer bestimmten Genauigkeit ermittelt werden können (Rundungsfehler, Diskretisierungsfehler, Abbruchfehler bei der Iteration usw.).

2.3.1 Darstellung von Funktionen und Operatoren

In unserem Verfahren werden "abgebrochene" Funktionen $\hat{x}$ im Intervall $[0, T]$ durch einen Satz von N äquidistanten Abtastwerten $\hat{x}(i\Delta T)$ mit $i=0, \dots, N-1$ näherungsweise dargestellt (siehe Abb. (3)). Dabei ist $\Delta T = T/(N-1)$.

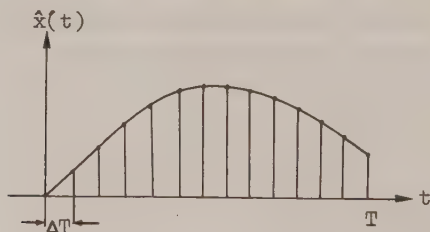


Abb.(3)

Da die betrachteten (quadratisch integrierbaren) Funktionen im allgemeinen nicht bandbegrenzt sind, geht man durch diese Diskretisierung einen Fehler im Sinne eines Informationsverlustes ein, der aber durch die Wahl eines hinreichend großen N beliebig klein gehalten werden kann. Damit kann man aus den Abtastwerten durch bandbegrenzte Interpolation eine Funktion finden, die die ursprüngliche Funktion hinreichend genau approximiert (im Sinne des Fehlerquadrats). In unserem Verfahren wird anstelle der bandbegrenzten Interpolation die numerisch einfachere Spline-Interpolation dritter Ordnung verwendet, da auch diese einen tiefpaßartigen Charakter aufweist.

Die Approximation einer vorgegebenen Funktion durch eine Spline-Interpolation dritter Ordnung über N Stützstellen liefert in der Nähe der Sprungstellen der Funktion einen relativ großen Fehler, der qualitativ dem Gibbsschen Phänomen entspricht (siehe Abb. (4)).

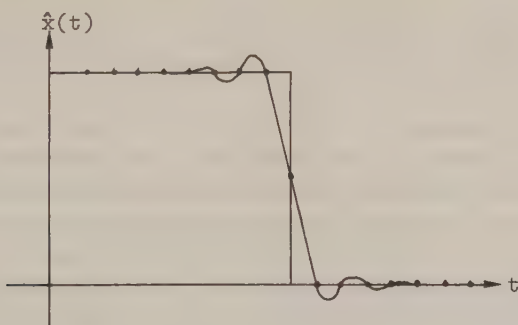


Abb. (4)

Der nichtlineare Operator $\mathbf{N} \in \Psi(\alpha, \beta)$ wird dadurch realisiert, daß aus der theoretischen Abbildung

$$y = \mathbf{N}x$$

gemäß

$$y(t) = \psi(x(t), t) \quad t \in [0, T]$$

die diskretisierte Abbildungsvorschrift

$$y(i\Delta T) = \psi(x(i\Delta T), i\Delta T) \quad \text{für } i=0, \dots, N-1$$

abgeleitet wird. Es ist klar, daß der Operator $\mathbf{N}$ nur an den Stützstellen genau realisiert wird, und deshalb muß man auch hier durch die Wahl eines geeignet großen N (bzw. kleinen ΔT) Sorge dafür tragen, daß der Fehler zwischen der exakten und der approximierenden Funktion hinreichend klein wird.

Alle linearen Abbildungen werden in unserem Verfahren mit dem Hilfsmittel der Fourier-Transformation durchgeführt. Zunächst entsteht für uns damit das Problem der numerischen Fourier-Transformation bzw. Rücktransformation von Funktionen. Dabei wollen wir nicht nur Zeitfunktionen, sondern auch Frequenzfunktionen (Real- bzw. Imaginärteil von Spektralfunktionen) als Spline-Interpolationen dritter Ordnung durch äquidistante Stützstellen gemäß Abb. (5) darstellen. Gaißmaier hat in [3] ein schnelles und genaues Verfahren vorgeschlagen, das unter Verwendung der schnellen Fourier-Transformation (FFT) N Stützstellen einer kausalen spline-interpolierten Zeitfunktion in N Stützstellen der Fourier-Transformierten dieser Zeitfunktion umrechnet. (Auf Grund der FFT muß dabei N eine Zweierpotenz sein.) Umgekehrt erlaubt das Verfahren die Berechnung von N Stützstellen einer als kausal vorausgesetzten Zeitfunktion, aus der durch N Stützstellen spline-interpolierten Realteilmfunktion einer vorgegebenen Spektralfunktion. Der Zeitabstand ΔT bzw. Kreisfrequenzabstand $\Delta \Omega$ zweier benachbarter Stützstellen ist dabei durch

$$\Delta T = \frac{2\pi}{\Omega}$$

bzw.

$$\Delta \Omega = \frac{2\pi}{T}$$

gegeben, wobei

$$\Delta T = T/(N-1)$$

und

$$\Delta \Omega = \Omega/(N-1)$$

gilt. T ist also die größte betrachtete Zeit, und Ω ist die größte betrachtete Kreisfrequenz. Da sowohl die Zeitunktionen als auch die Spektralfunktionen "abgebrochen" und beschränkt sind, können die L_2 -Fourier-Transformierten bzw. Rücktransformierten als gewöhnliche Integrale berechnet werden (siehe Bemerkung 2 von Abschnitt I.4.2).

Abgesehen von Rundungsfehlern treten bei diesem Verfahren zwei systematische Fehlerquellen auf:

- (F 1) Die Zeit- bzw. Frequenzfunktionen lassen sich nur mit einer gewissen Genauigkeit durch Spline-Funktionen annähern. (N muß also bei vorgegebenem T bzw. Ω hinreichend groß gewählt werden.)
- (F 2) Bei der Rücktransformation einer vorgegebenen Spektralfunktion werden Spektralanteile, die außerhalb der maximalen Kreisfrequenz Ω liegen, vernachlässigt. (Ω muß also hinreichend groß gewählt werden.)

Die Anwendung des linearen Operators $\mathbf{F}$ auf eine "abgebrochene" Funktion $\hat{x} \in L_{2r}^n(-\infty, \infty)$ sei im Frequenzbereich mit $y = \mathbf{F}\hat{x}$ durch

$$Y(j\omega) = \mathbf{F}(j\omega)X(j\omega)$$

für alle $\omega \in \mathbf{R}$ definiert. Daraus leiten wir für unser Verfahren die diskretisierte Abbildungsvorschrift

$$Y(j(i\Delta\Omega)) = \mathbf{F}(j(i\Delta\Omega))X(j(i\Delta\Omega))$$

für $i=0, \dots, N-1$ ab, wobei $X(j(i\Delta\Omega))$ bzw. $Y(j(i\Delta\Omega))$ der i -te Abtastwert der Fourier-Transformierten von $\hat{x}$ bzw. y ist. Die Werte $F(j(i\Delta\Omega))$ für $i=0, \dots, N-1$ können aus der Übertragungsfunktion des linearen Systems berechnet werden, während die numerische Fourier-Transformation von $\hat{x}(t)$ bzw. Rücktransformation von $Y(j\omega)$ durch das Verfahren von Gaißmaier erfolgen soll. (Da die Stützstellen von $y(t)$ nur bis zum Zeitpunkt T berechnet werden, wird damit der Operator P_{TF} näherungsweise realisiert.)

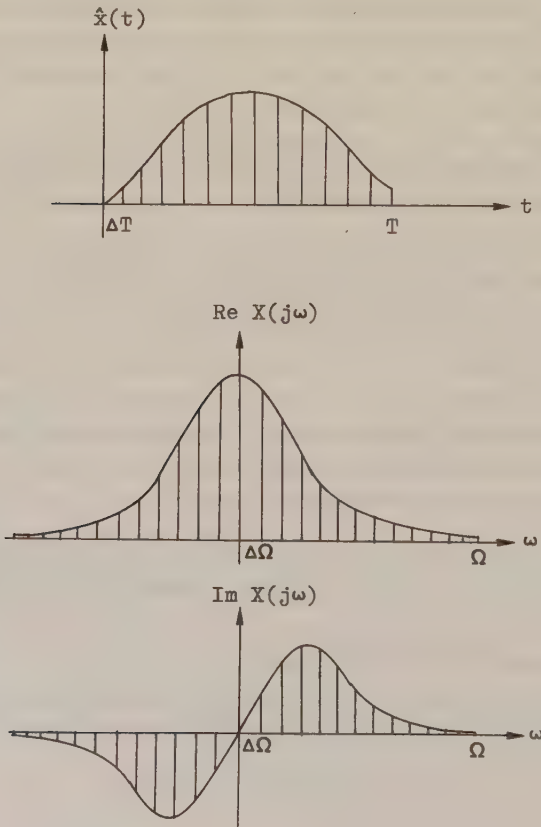


Abb. (5)

2.3.2 Das Lösungsverfahren

Wir wollen jetzt die praktische Berechnung der bei der Maximalzeit T "abgebrochenen" Funktionen $\hat{e}$, $\hat{i}$ und $\hat{r}$ gemäß Satz 2 von Abschnitt IV.2.2 beschreiben. Bei dieser Berechnung werden alle Funktionen und Operatoren wie in Abschnitt IV.2.3.1 dargestellt. Dabei werden aus Gründen der Vereinfachung Näherungs- und exakte Werte gleich bezeichnet, und wie immer werden durch Kleinbuchstaben Zeitfunktionen und durch die entsprechenden Großbuchstaben Frequenzfunktionen dargestellt.

Berechnungsschritt (a):

Theorie:

$$\hat{e}_0 = \mathbf{P}_T \mathbf{e}_0$$

Verfahren:

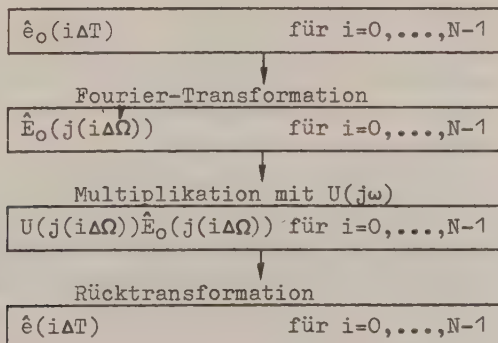
$\hat{e}_0(i\Delta T) = e_0(i\Delta T) \quad \text{für } i=0, \dots, N-1$

Berechnungsschritt (b):

Theorie:

$$\hat{e} = \mathbf{P}_T \mathbf{U} \hat{e}_0$$

Verfahren:



Berechnungsschritt (c):

Theorie:

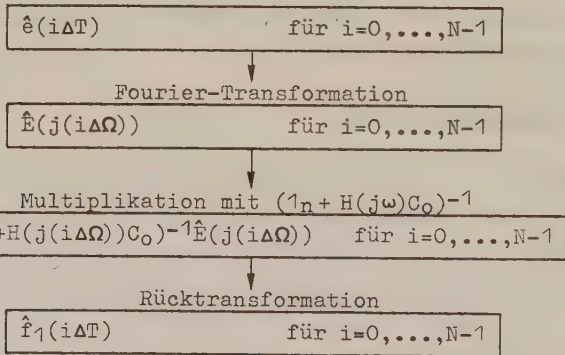
$$\hat{f} = \lim_{\mu \rightarrow \infty} \hat{f}_{\mu}$$

mit

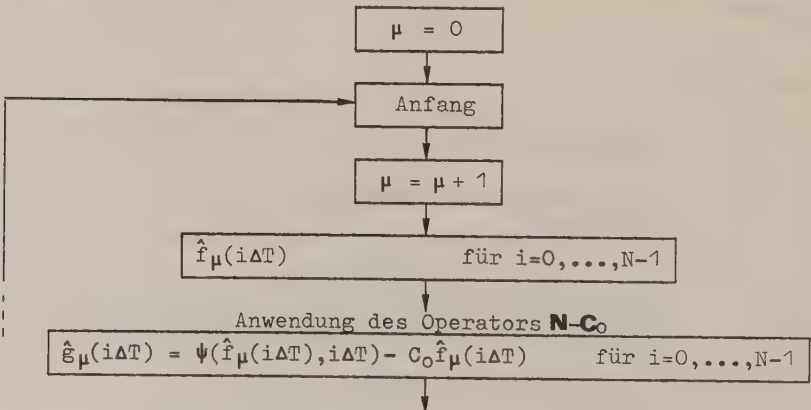
$$\hat{f}_{\mu+1} = -\mathbf{P}_T(\mathbf{I} + \mathbf{H}\mathbf{C}_0)^{-1}\mathbf{H}(\mathbf{N}-\mathbf{C}_0)\hat{f}_{\mu} + \mathbf{P}_T(\mathbf{I} + \mathbf{H}\mathbf{C}_0)^{-1}\hat{e}$$

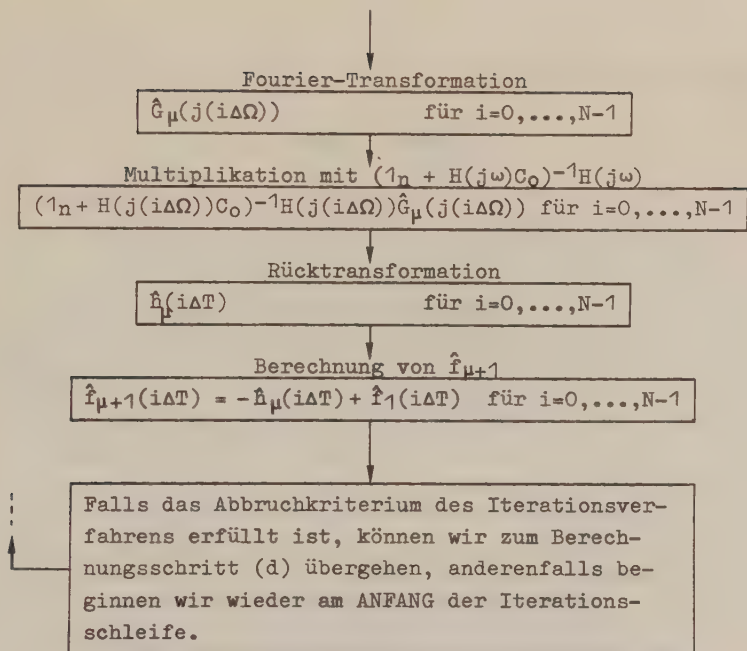
für $\mu=0,1,\dots$ und $\hat{f}_0 = \theta$

Verfahren:



Jetzt folgt das Iterationsverfahren:





Das Abbruchkriterium ist erfüllt, wenn

- (A1) keine Konvergenz des Verfahrens (die Zahl der Iterationsschritte ist größer als die zulässige Schrittzahl) oder
- (A2) der Abstand zwischen den letzten beiden Funktionen $\hat{f}_{\mu+1}$ und $\hat{f}_\mu$ ist größer als der Abstand zwischen $\hat{f}_\mu$ und $\hat{f}_{\mu-1}$. Anschaulich bedeutet dies, daß das Verfahren nur noch Fehler berechnet (Rundungsfehler), da ja wegen

$$\|f_{\mu+1} - f_\mu\| \leq r_0 \|f_\mu - f_{\mu-1}\|$$

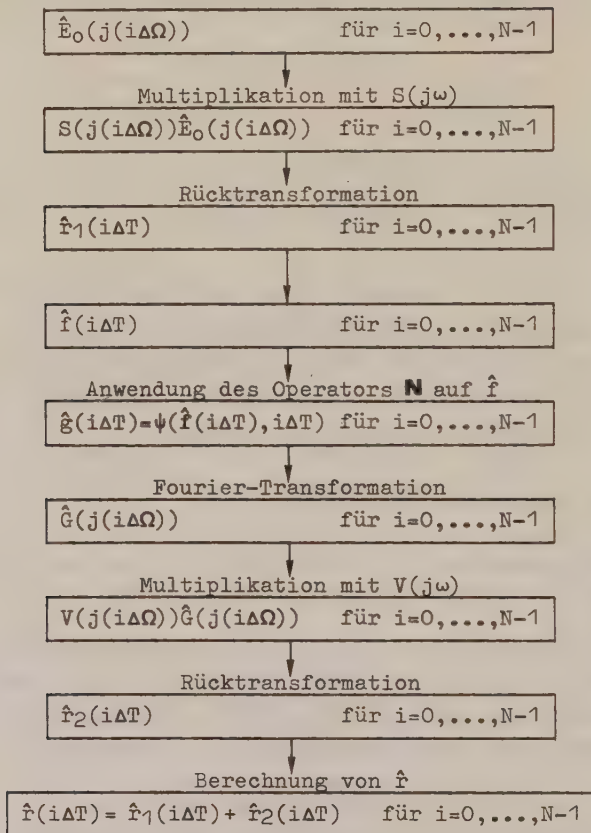
der Abstand immer kleiner werden müßte, und daher dieser Fall nur von Berechnungsfehlern herrühren kann. $\hat{f}_\mu$ stellt also die beste Näherung dar, die das Verfahren zu liefern imstande ist (siehe auch Abschnitt IV.2.3.3).

Berechnungsschritt (d):

Theorie:

$$\hat{\mathbf{r}} = \mathbf{P}_T \mathbf{S} \hat{\mathbf{e}}_0 + \mathbf{P}_T \mathbf{V} \mathbf{N} \hat{\mathbf{f}}$$

Verfahren:



Damit ist die gesuchte Systemreaktion $\hat{\mathbf{r}}$ näherungsweise gefunden.

Abschließend beschreiben wir noch die numerische Überprüfung des Konvergenzkriteriums. Es war

$$r_0 = \sup_{\omega \in \mathbf{R}} \Lambda \{ (1_n + H(j\omega)C_0)^{-1} H(j\omega)T_0^{-1} \} \eta(C_0, T_0)$$

mit

$$\eta(C_0, T_0) = \max_{i=1, \dots, n} (t_i^{(0)} \eta_i(c_i^{(0)})) = \frac{1}{2} \max_{i=1, \dots, n} (t_i^{(0)} (\beta_i - \alpha_i)).$$

Die einzige Schwierigkeit besteht in der Berechnung des Supremums. In unserem Verfahren wird dieses durch eine Maximumbildung über N Stützstellen angenähert, d.h.

$$\begin{aligned} & \sup_{\omega \in \mathbf{R}} \Lambda \{ (1_n + H(j\omega)C_0)^{-1} H(j\omega)T_0^{-1} \} \\ & \approx \max_{i=0, \dots, N-1} \Lambda \{ (1_n + H(j(i\Delta\Omega)C_0)^{-1} H(j(i\Delta\Omega)T_0^{-1} \}, \end{aligned}$$

da $\Lambda \{ (1_n + H(j(\omega)C_0)^{-1} H(j(\omega)T_0^{-1} \}$ in ω stetig ist. Dabei muß durch geeignete Wahl von N und Ω sichergestellt sein, daß diese Relation sinnvoll ist.

2.3.3 Konvergenz und Genauigkeit

Theoretische Aussagen können durch numerische Verfahren nur bis auf eine gewisse Genauigkeit, die von Rundungsfehlern und systematischen Fehlern abhängt, umgesetzt werden.

In unserem Verfahren wird die Iterationsvorschrift

$$\hat{f}_{\mu+1} = \Phi \hat{f}_{\mu} + \hat{f}_1 \quad \text{für } \mu=1, 2, \dots \quad (2)$$

durch die Iterationsvorschrift

$$\hat{f}_{\mu+1}^N = \Phi_N \hat{f}_{\mu}^N + \hat{f}_1^N \quad \text{für } \mu=1, 2, \dots \quad (3)$$

ersetzt, wobei Φ_N den gegebenen Operator Φ auf einer Rechenanlage gemäß Abschnitt IV.2.3.1 bzw. IV.2.3.2 näherungsweise darstellt. Dabei ist $g_0(\Phi) < 1$, und wir setzen zur Vereinfachung $\hat{f}_1 = \hat{f}_1^N$. (Wir kümmern uns hier also nur um den Einfluß des Fehlers von Φ , da Fehler, die bei der Darstellung der anderen Operatoren entstehen, bereits in Abschnitt IV.2.3.1 diskutiert wurden.) Obwohl wir wissen, daß die Iterationsvorschrift gemäß Gl. (2) konvergiert, ist es zunächst nicht sicher, daß unser numerisches Iterationsverfahren nach Gl. (3) befriedigend arbeitet. Im folgenden untersuchen wir, inwieweit die Rundungsfehler und systematischen Fehler das Konvergenzverhalten unseres Verfahrens beeinflussen.

Es wäre schön, wenn wir zeigen könnten, daß Φ_N eine Kontraktion ist. Das ist aber nicht zu erwarten, da bei Anwendung von Φ_N auf irgendeine Funktion x der Abbildungsfehler Δy gemäß

$$\Delta y = \Phi_N x - \Phi x$$

im allgemeinen in keiner einfachen Beziehung zu x steht und deshalb die Zahl

$$\|\Phi_N x_1 - \Phi_N x_2\| = \|\Phi x_1 - \Phi x_2 + \Delta y_1 - \Delta y_2\|$$

nicht durch $\|x_1 - x_2\|$ abgeschätzt werden kann. (Die Abschätzung müßte ja für alle zulässigen x_1, x_2 und insbesondere für $x_1 \approx x_2$ gelten.) Der Operator Φ_N wirkt aber global wie eine Kontraktion, wenn mit

$$\begin{aligned} \|\Phi_N x_1 - \Phi_N x_2\| &= \|(\Phi x_1 - \Phi x_2) + (\Delta y_1 - \Delta y_2)\| \\ &\leq r_0 \|x_1 - x_2\| + \|\Delta y_1 - \Delta y_2\| \\ &= \left(r_0 + \frac{\|\Delta y_1 - \Delta y_2\|}{\|x_1 - x_2\|}\right) \|x_1 - x_2\| \end{aligned}$$

die Ungleichung

$$\frac{\|\Delta y_1 - \Delta y_2\|}{\|x_1 - x_2\|} < 1 - r_0 \quad (4)$$

gilt, d.h. er besitzt nur für hinreichend weit auseinanderliegende x_1, x_2 die Kontraktionseigenschaft. Folglich ist klar, daß

$$\|\hat{f}_{\mu+1}^N - \hat{f}_{\mu}^N\| < \|\hat{f}_{\mu}^N - \hat{f}_{\mu-1}^N\|$$

gilt und damit die "Konvergenzeigenschaft" der Folge erhalten bleibt, wenn $\hat{f}_{\mu}^N$ und $\hat{f}_{\mu-1}^N$ im Sinne von Gl. (4) hinreichend weit auseinander liegen. Die Folge $\{\hat{f}_{\mu}^N\}$ wird aber im streng mathematischen Sinne nicht konvergieren, da Φ_N lokal keine Kontraktion ist, und wir bekommen also grundsätzlich den in Abb.(6) dargestellten Verlauf von $\|\hat{f}_{\mu+1}^N - \hat{f}_{\mu}^N\|$ in Abhängigkeit von μ .

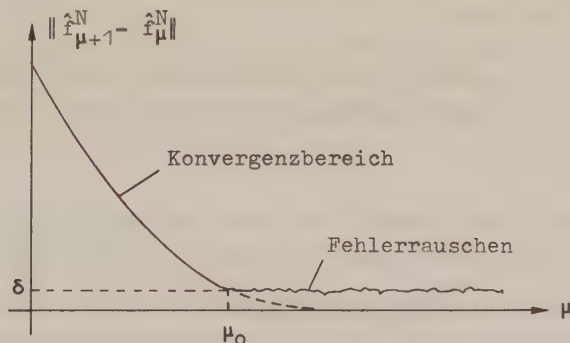


Abb. (6)

Das bedeutet in der Praxis, daß ab einer gewissen Zahl μ_0 von Iterationsschritten der Abstand zwischen $\hat{f}_{\mu+1}^N$ und $\hat{f}_{\mu}^N$ nicht mehr kleiner gemacht werden kann. Ein sinnvolles Kriterium für den Abbruch der Iteration ist deshalb

$$\|\hat{f}_{\mu_0+1}^N - \hat{f}_{\mu_0}^N\| > \|\hat{f}_{\mu_0}^N - \hat{f}_{\mu_0-1}^N\|$$

und man wird dann $\hat{f}_{\mu_0}^N$ als beste Näherungslösung von Gl. (2) ansehen. Es sei noch erwähnt, daß der Abstand δ der letzten beiden Iterationsschritte mit den letzten Berechnungsfehlern

$$\Delta \hat{f}_{\mu_0}^N = \Phi_N \hat{f}_{\mu_0}^N - \Phi \hat{f}_{\mu_0}^N$$

und

$$\Delta \hat{f}_{\mu_0-1}^N = \Phi_N \hat{f}_{\mu_0-1}^N - \Phi \hat{f}_{\mu_0-1}^N$$

gemäß

$$\delta = \|\hat{f}_{\mu_0}^N - \hat{f}_{\mu_0-1}^N\| \approx \frac{1}{1-r_0} \|\Delta \hat{f}_{\mu_0}^N - \Delta \hat{f}_{\mu_0-1}^N\|$$

zusammenhängt, weil dadurch der Übergang zwischen konvergentem und divergentem Verhalten der Iterationsvorschrift definiert ist. (Diese Relation entsteht aus Gl. (4), indem man $<$ durch $\approx$ ersetzt.)

Als letztes wichtiges Problem bleibt zu klären, inwieweit $\hat{f}_{\mu_0}^N$ als gute Näherungslösung der Gleichung

$$\hat{f} = \Phi \hat{f} + \hat{f}_1 \quad (5)$$

aufgefaßt werden kann. Dazu subtrahieren wir von Gl. (5) die Gleichung des letzten Iterationsschrittes

$$\hat{f}_{\mu_0}^N = \Phi_N \hat{f}_{\mu_0}^N + \hat{f}_1^N$$

und erhalten wegen $\hat{f}_1 = \hat{f}_1^N$

$$\hat{f} - \hat{f}_{\mu_0}^N = \Phi \hat{f} - \Phi_N \hat{f}_{\mu_0}^N$$

bzw.

$$\hat{f} - \hat{f}_{\mu_0}^N = \Phi \hat{f} - \Phi \hat{f}_{\mu_0-1}^N - \Delta \hat{f}_{\mu_0-1}^N.$$

Da $g_0(\Phi) \leq r_0 < 1$ ist, folgt

$$\begin{aligned}
 \|\hat{f} - \hat{f}_{\mu_0}^N\| &\leq \|\Phi\hat{f} - \Phi\hat{f}_{\mu_0-1}^N\| + \|\Delta\hat{f}_{\mu_0-1}^N\| \\
 &\leq r_0 \|\hat{f} - \hat{f}_{\mu_0-1}^N\| + \|\Delta\hat{f}_{\mu_0-1}^N\| \\
 &\leq r_0 (\|\hat{f} - \hat{f}_{\mu_0}^N\| + \|\hat{f}_{\mu_0}^N - \hat{f}_{\mu_0-1}^N\|) + \|\Delta\hat{f}_{\mu_0-1}^N\|
 \end{aligned}$$

und schließlich

$$\|\hat{f} - \hat{f}_{\mu_0}^N\| \leq \frac{1}{1-r_0} \left\{ r_0 \|\hat{f}_{\mu_0}^N - \hat{f}_{\mu_0-1}^N\| + \|\Delta\hat{f}_{\mu_0-1}^N\| \right\}.$$

Anhand dieser Gleichung sind wir in der Lage, den Fehler der Näherungslösung $\hat{f}_{\mu_0}^N$ durch den Abstand der letzten beiden Näherungen und durch den Fehler $\|\Delta\hat{f}_{\mu_0-1}^N\|$ abzuschätzen. Dabei ist $\|\Delta\hat{f}_{\mu_0-1}^N\|$ der Fehler, der dadurch entsteht, daß nicht der ideale Operator Φ , sonder der reale Operator Φ_N auf die Funktion $\hat{f}_{\mu_0-1}^N$ angewandt wird. Beide Fehlerterme entstehen letztlich auf Grund der Rundungsfehler und systematischen Fehler des Verfahrens, und sie können durch hinreichend feine Diskretisierung und große Rechengenauigkeit beliebig klein gemacht werden.

3. Analysebeispiele

Das in Abschnitt IV.2 beschriebene Analyseverfahren wurde an einer Reihe praktischer Beispiele erfolgreich erprobt. Bei all diesen Beispielen wird es sich zeigen, daß die theoretischen Aussagen von Kapitel II. in einem gewissen Sinne scharfe Aussagen sind. Obwohl die Konvergenz-, Existenz- und Eindeutigkeitskriterien hinreichende Kriterien sind, zeigt es sich, daß bei einer Verletzung dieser Bedingungen häufig auch die Aussagen verletzt werden. Ebenso stellen die Abschätzungen bezüglich der Konvergenzgeschwindigkeit des Verfahrens keine "akademischen" Aussagen dar, sondern sie be-

schreiben das tatsächliche Konvergenzverhalten meistens sehr gut. In diesem Sinne sind die Kriterien für die Anwendbarkeit des Verfahrens nicht nur hinreichend, sondern "fast" notwendig.

Die notwendigen numerischen Rechnungen wurden an den Rechenanlagen CD 3300 und TR 440 der Universität Erlangen-Nürnberg durchgeführt.

3.1 Systeme mit konzentrierten Elementen

3.1.1 Beispiel K1

Gegeben sei das Rückkopplungssystem der Abb. (7) mit der skalaren Erregung e gemäß Abb. (8) und der skalaren Systemreaktion f .

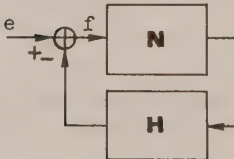


Abb. (7)

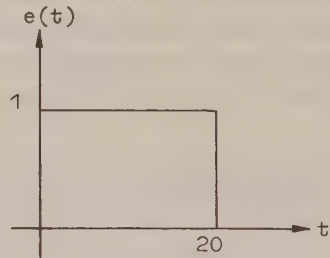


Abb. (8)

Der lineare Operator **H** sei durch die Übertragungsfunktion

$$H(p) = \frac{2}{p^2 + 0.6p + 1}$$

und der nichtlineare Operator **N** durch die nichtlineare Funktion $\psi(x)$ gemäß Abb. (9) definiert. Dabei ist $\alpha=0.2$ und $\beta=1.0$.

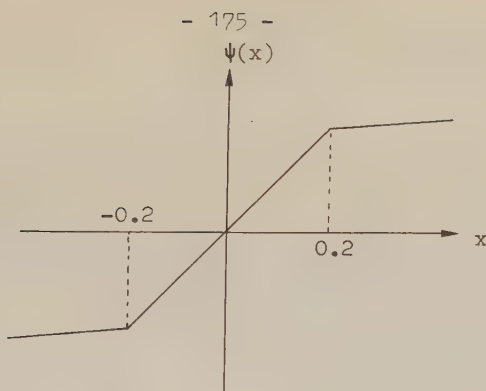


Abb. (9)

Aus dem Frequenzgang der Abb. (10) ist ersichtlich, daß für diese Nichtlinearität das Frequenzgangkriterium gemäß Abschnitt II.2.5.3.2 erfüllt ist, und damit existiert eine eindeutige Systemreaktion f . In Abb. (11) sind die Funktionen $\hat{f}_\mu$ für $\mu=1, \dots, 15$ eingezeichnet. Die Konvergenzgeschwindigkeit entspricht dabei in etwa der Kontraktionsschranke $r_0 = 0.92$. In Abb. (12) wird das Iterationsverfahren bei einer nichtoptimalen Linearisierung demonstriert. Die Kontraktionsschranke beträgt hier $r = 1.56$, und somit ist die Konvergenz des Verfahrens nicht gesichert. Daß tatsächlich Divergenz auftritt, ist ein Zeichen dafür, daß das Konvergenzkriterium "fast" notwendig ist. Abb. (13) zeigt das Verhalten des Iterationsprozesses beim System der Abb. (7), wobei jetzt der Knick der nichtlinearen Kennlinie erst bei $x = 0.5$ auftritt.

FREQUENZGANG VON BEISPIEL NR. K1

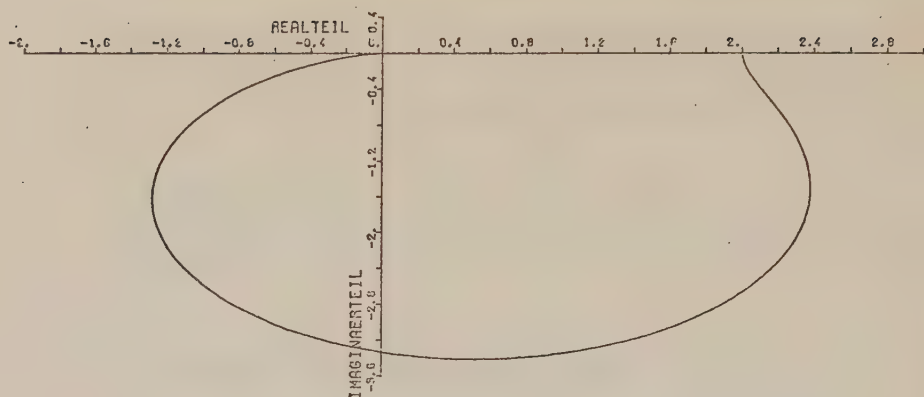


Abb.(10)

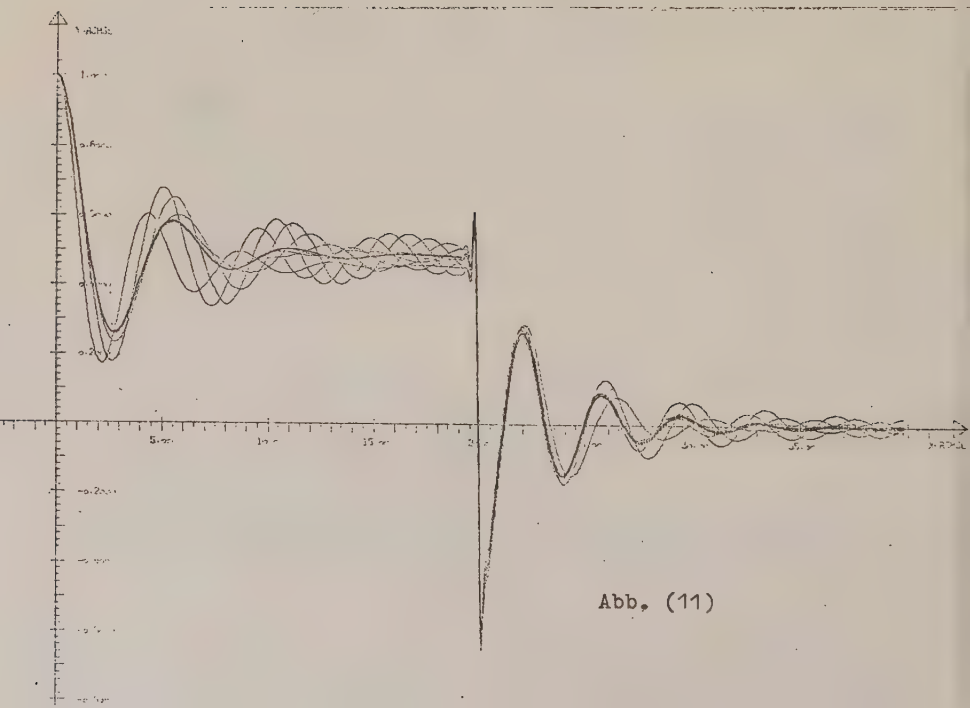


Abb. (11)

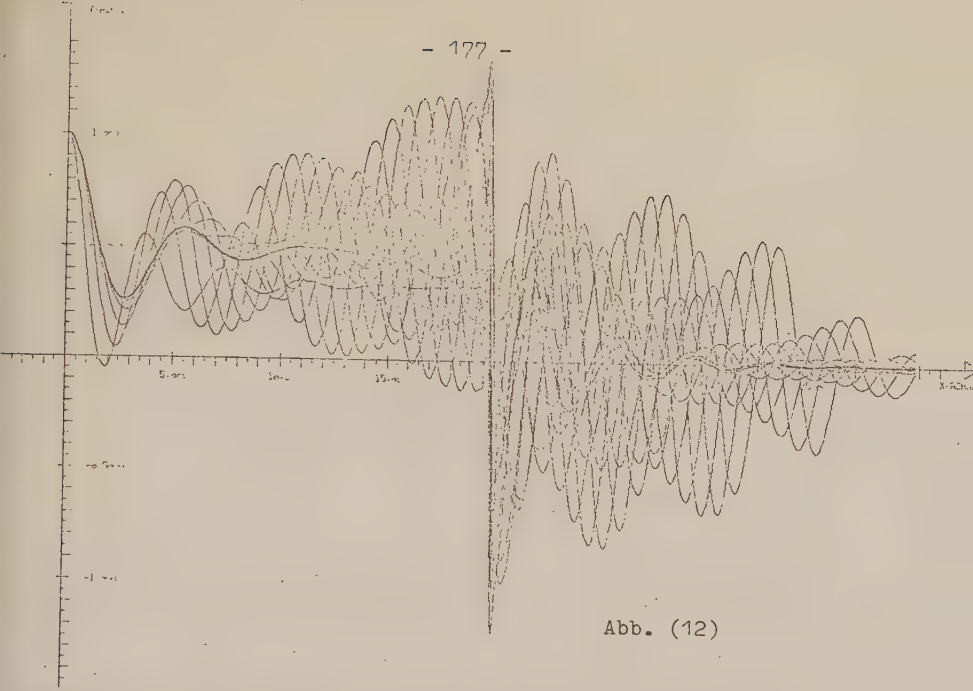


Abb. (12)

ITERATION VON BEISPIEL NR. K1 MIT $A=0.2$ $B=1.0$ ($\epsilon=1.0$)

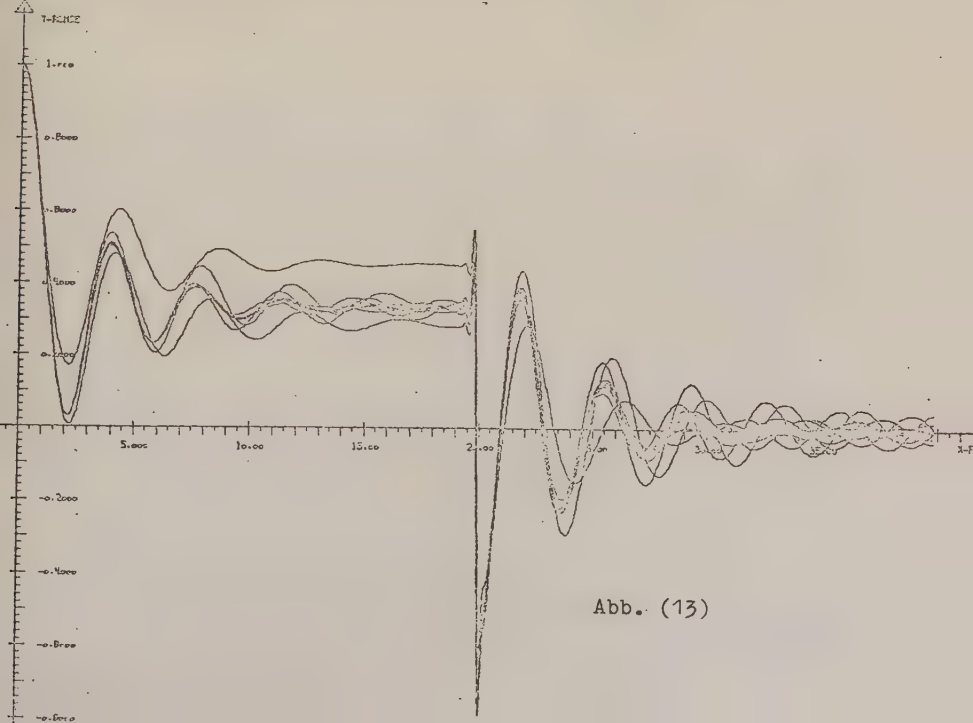


Abb. (13)

3.1.2 Beispiel K2

Gegeben sei das Rückkopplungssystem der Abb. (14) mit der skalaren Erregung e gemäß Abb. (17) und der skalaren Systemreaktion r . Die linearen Operatoren $\mathbf{H}$ bzw. $\mathbf{V}$ seien durch die Übertragungsfunktionen

$$H(p) = \frac{p^2 + 3.8p + 2.4}{p^5 + 7.2p^4 + 16.4p^3 + 17.8p^2 + 15.6p + 2.}$$

bzw.

$$V(p) = \frac{1}{p^2 + 0.3p + 2.}$$

und der nichtlineare Operator $\mathbf{N}$ durch die nichtlineare Funktion $\psi(x)$ gemäß Abb. (15) definiert. Dabei ist $\alpha = 0.$ und $\beta = 2.$

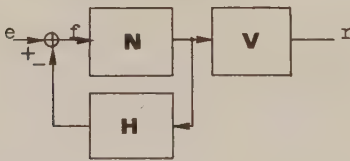


Abb. (14)

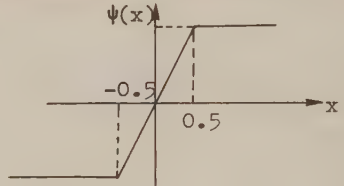
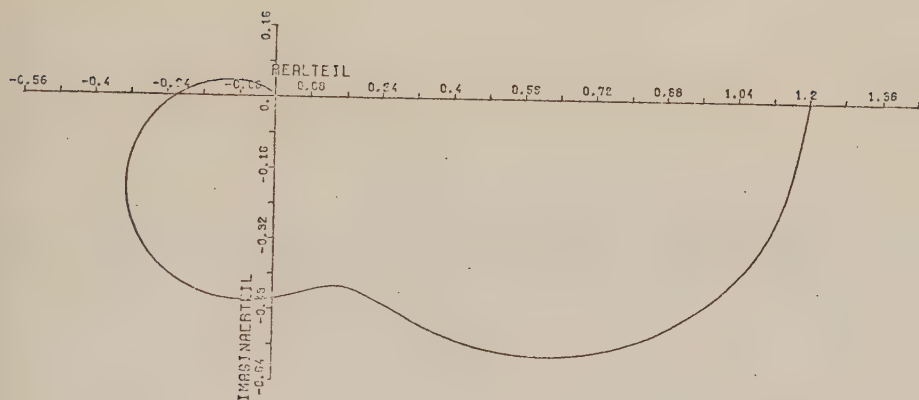


Abb. (15)

Aus dem Frequenzgang der Abb. (16) ist ersichtlich, daß für diese Nichtlinearität das Frequenzgangkriterium gemäß Abschnitt II.2.5.3.2 erfüllt ist, und damit existiert eine eindeutige Systemfunktion f . Die Kontraktionsschranke beträgt $r_0 = 0.57$. Die Systemfunktion $\hat{f}_1$ des durch $\mathbf{N}=\mathbf{C}_0$ linearisierten Systems zeigt Abb. (18), während Abb. (19) die Funktionen $\hat{f}_\mu$ für $\mu=1, \dots, \mu_0$ darstellt. Um die Genauigkeit des Verfahrens beurteilen zu können, werden in Abb. (20) die Funktionen $\hat{f}_{\mu_0}$ und $\hat{f}_{\mu_0-1}$ gemäß Abschnitt IV.2.3.3 gezeigt. Daß diese Funktionen visuell nicht voneinander unterscheidbar sind, bedeutet praktisch, daß die Rundungsfehler des Verfahrens unbedeutend sind. Als einzig relevante Fehlerquelle bleiben die systematischen Fehler des Verfahrens (Bandbegrenzung, Spline-Interpolation usw.). In den Abb. (21) und (22) sind schließlich die Systemreaktion des linearisierten Systems und des nichtlinearen Systems dargestellt.

FREQUENZGANG VON BEISPIEL NR. K2



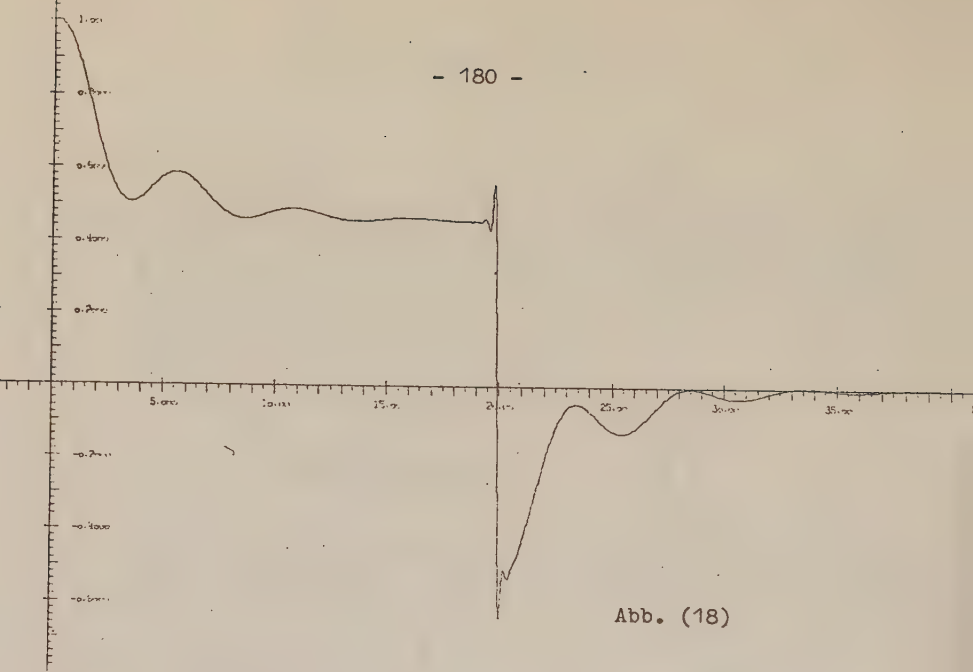


Abb. (18)

STE APPROXIMATION VON BEISPIEL NR. K2

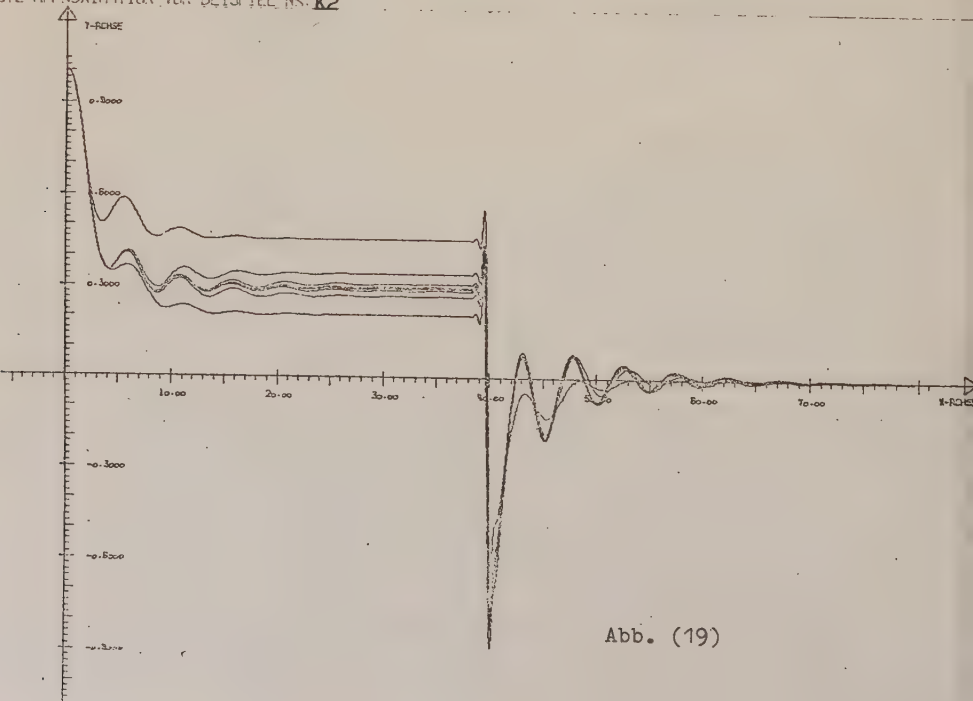


Abb. (19)

ITERATION VON BEISPIEL NR. K2 MIT $B=0$ $B=2$ $C=1$

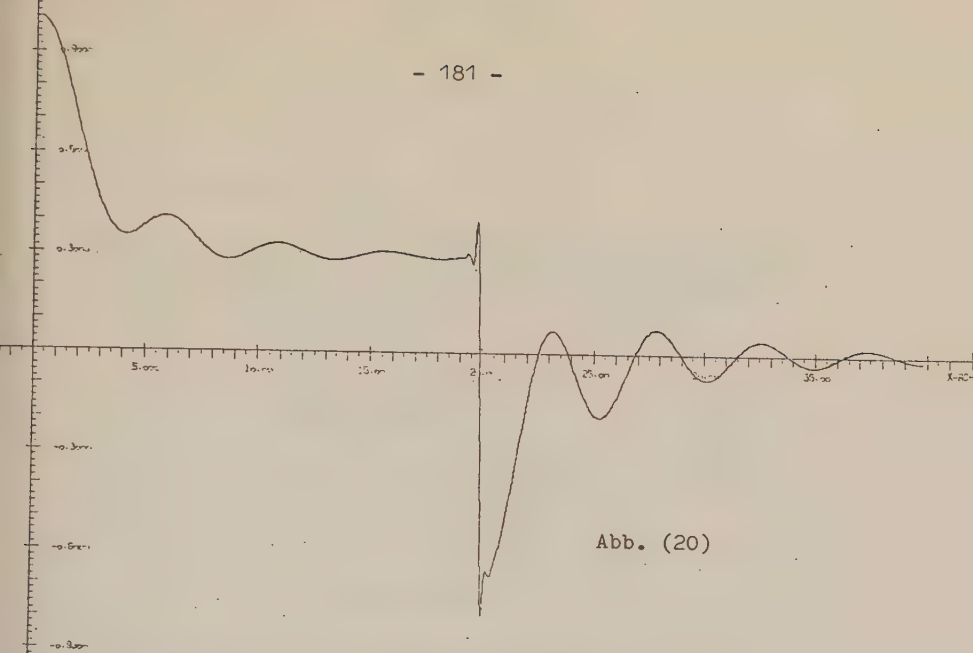


Abb. (20)

BEI DEN BEIDEN APPROXIMATIONEN VON BEISPIEL NR. K2

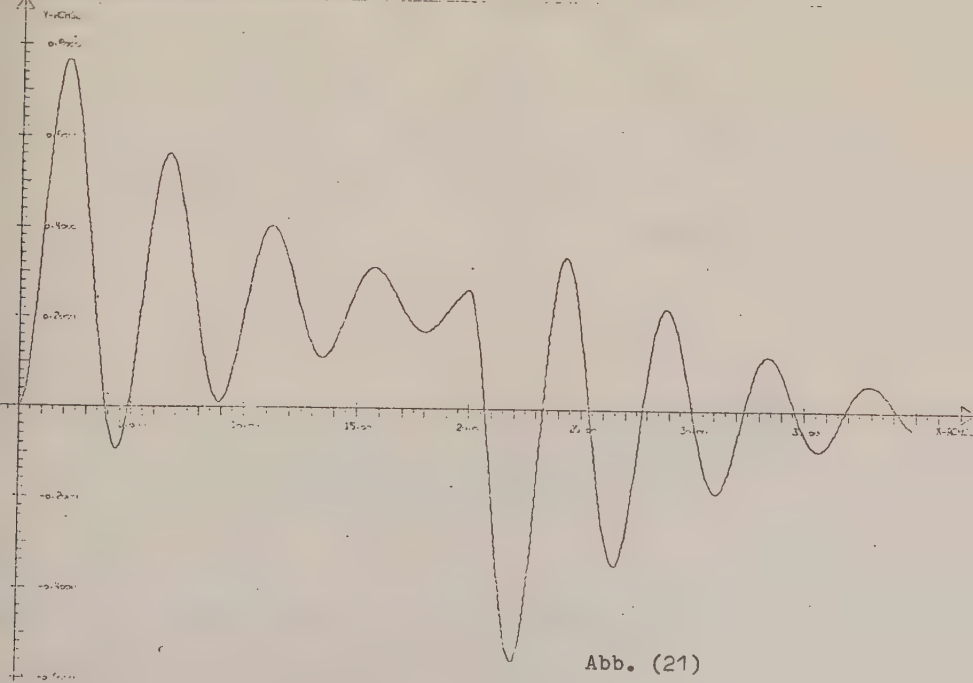
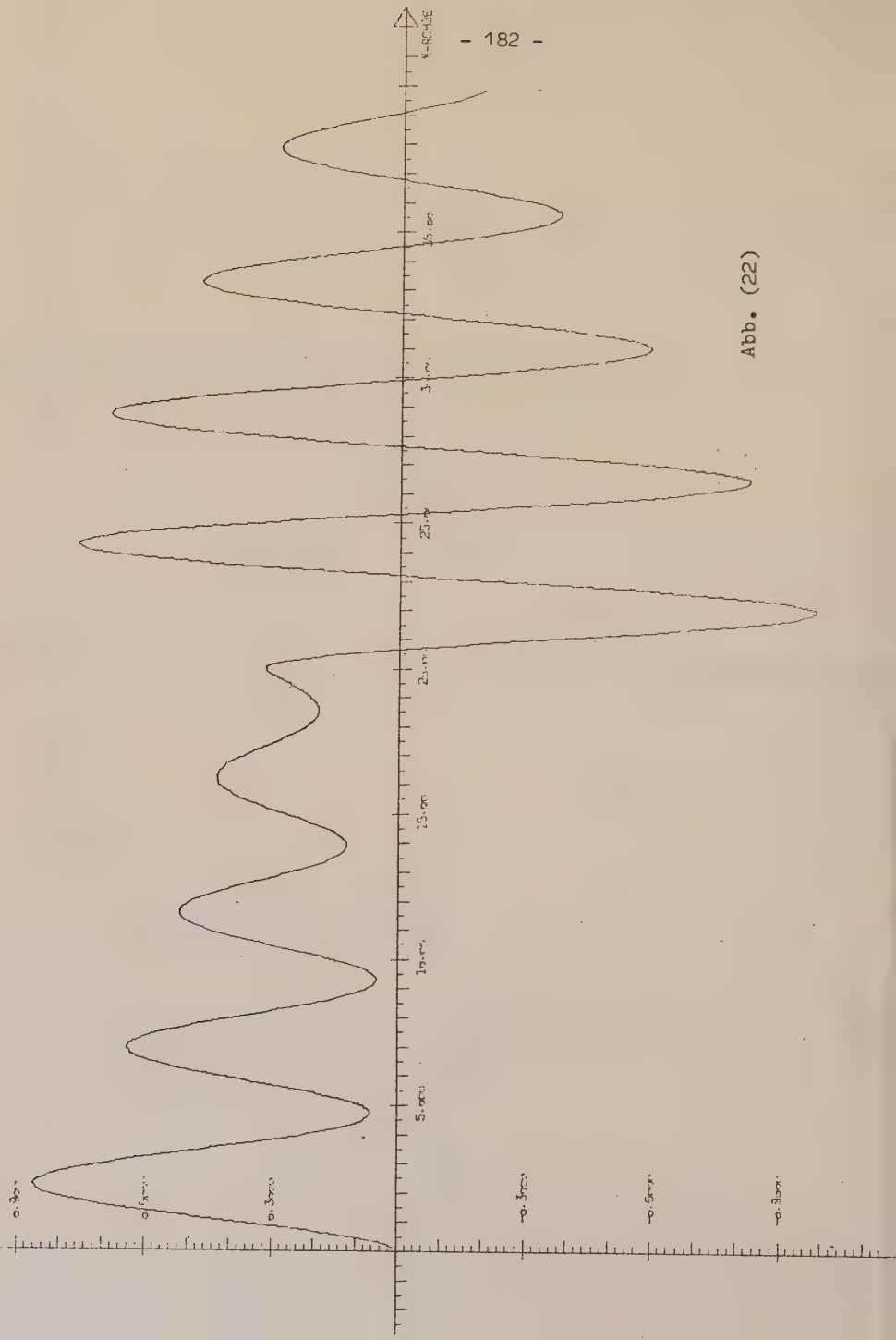


Abb. (21)



- 182 -

Abb. (22)

3.1.3 Beispiel K3

Gegeben sei die Gleichrichterschaltung der Abb. (23) mit der sinusförmigen (aber abgebrochenen) Erregungsspannung u_1 (siehe Abb. (26)) und der Ausgangsspannung u_2 .

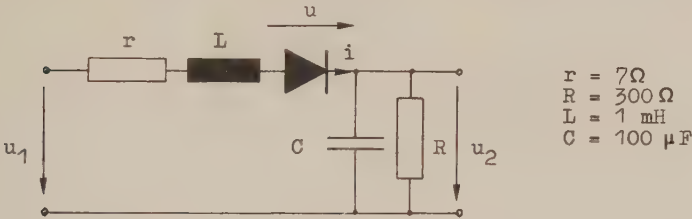


Abb. (23)

Gemäß Abschnitt III.3.1 kann diese Schaltung durch das Blockschaltbild der Abb. (24) dargestellt werden.

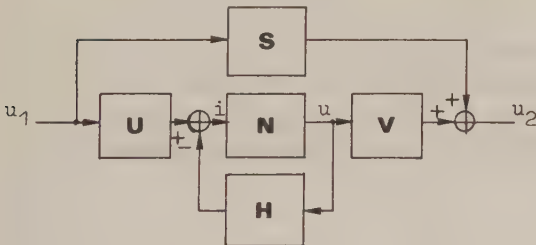


Abb. (24)

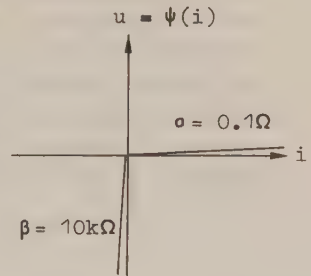


Abb. (25)

Der nichtlineare Operator **N** sei dabei durch die nichtlineare Funktion $\psi(i)$ gemäß Abb. (25) (α ist also der Durchlaß- und β der Sperrwiderstand der Diode) und die linearen Operatoren durch die Übertragungsfunktionen

$$U(p) = \frac{U_1(p)}{D(p)}, S(p) = \frac{S_1(p)}{D(p)}, H(p) = \frac{H_1(p)}{D(p)}, V(p) = \frac{V_1(p)}{D(p)}$$

dargestellt. Eine einfache lineare Analyse ergibt

mit $\hat{p} = p \cdot \text{sec}$:

$$D(p) = \hat{p}^2 + 0.703 \cdot 10^4 \hat{p} + 0.1023 \cdot 10^8$$

$$U_1(p) = (10^3 \hat{p} + 0.3 \cdot 10^5) S$$

$$H_1(p) = (10^3 \hat{p} + 0.3 \cdot 10^5) S$$

$$S_1(p) = 10^7$$

$$V_1(p) = -10^7.$$

Wegen Bemerkung 2 von Abschnitt III.3.1 kann unser Verfahren im Prinzip angewandt werden. Auf Grund des großen Unterschieds zwischen α und β beträgt die Kontraktionsschranke jedoch $r_0 = 0.9986$, und damit ist das Verfahren auf Grund der unvermeidlichen Berechnungsfehler hier unbrauchbar. Abb. (27) zeigt die letzten beiden Iterationsfunktionen für den Diodenstrom i , und Abb. (28) stellt die Ausgangsspannung u_2 dar. (Es waren 50 Iterationsschritte notwendig.)

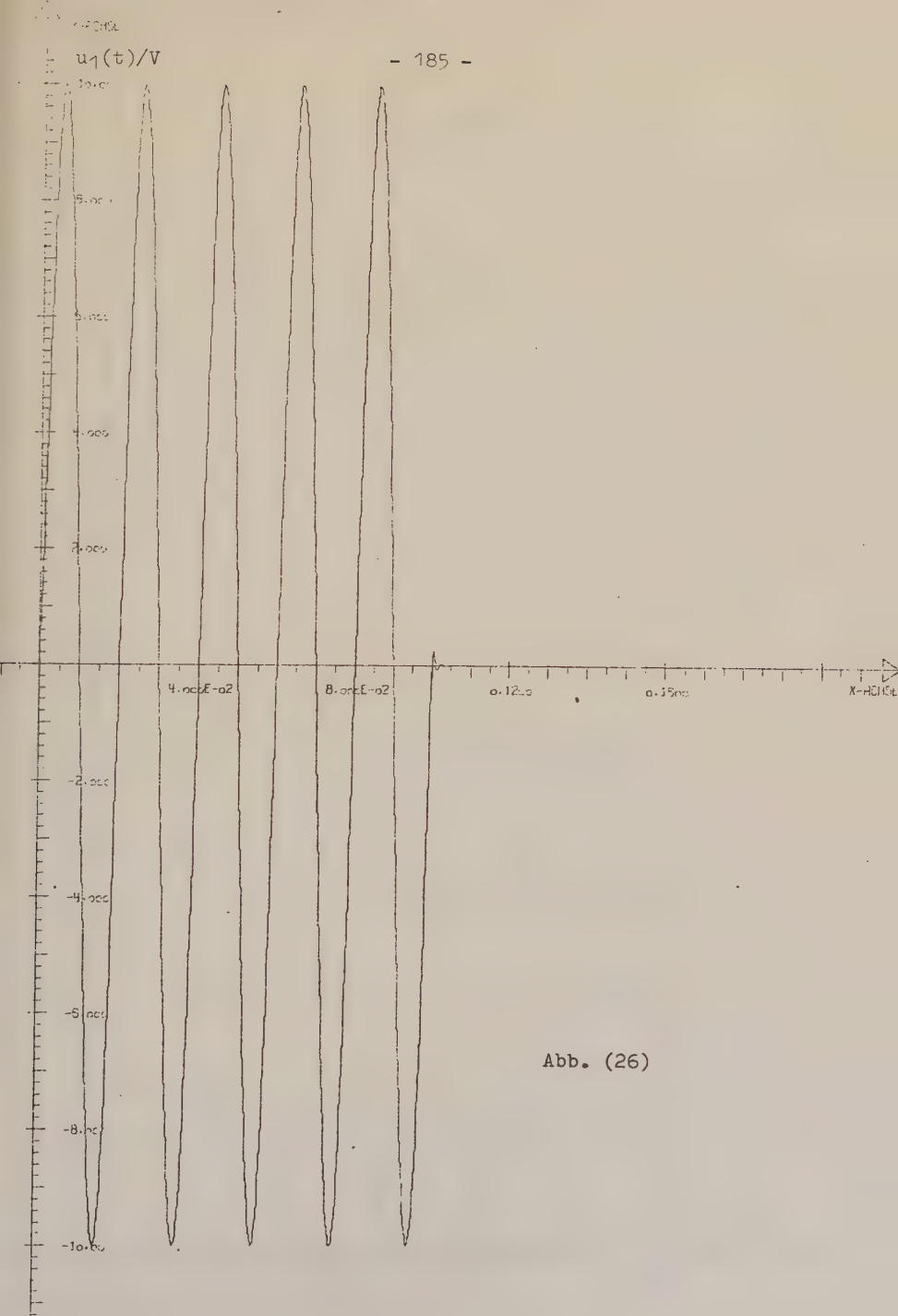


Abb. (26)

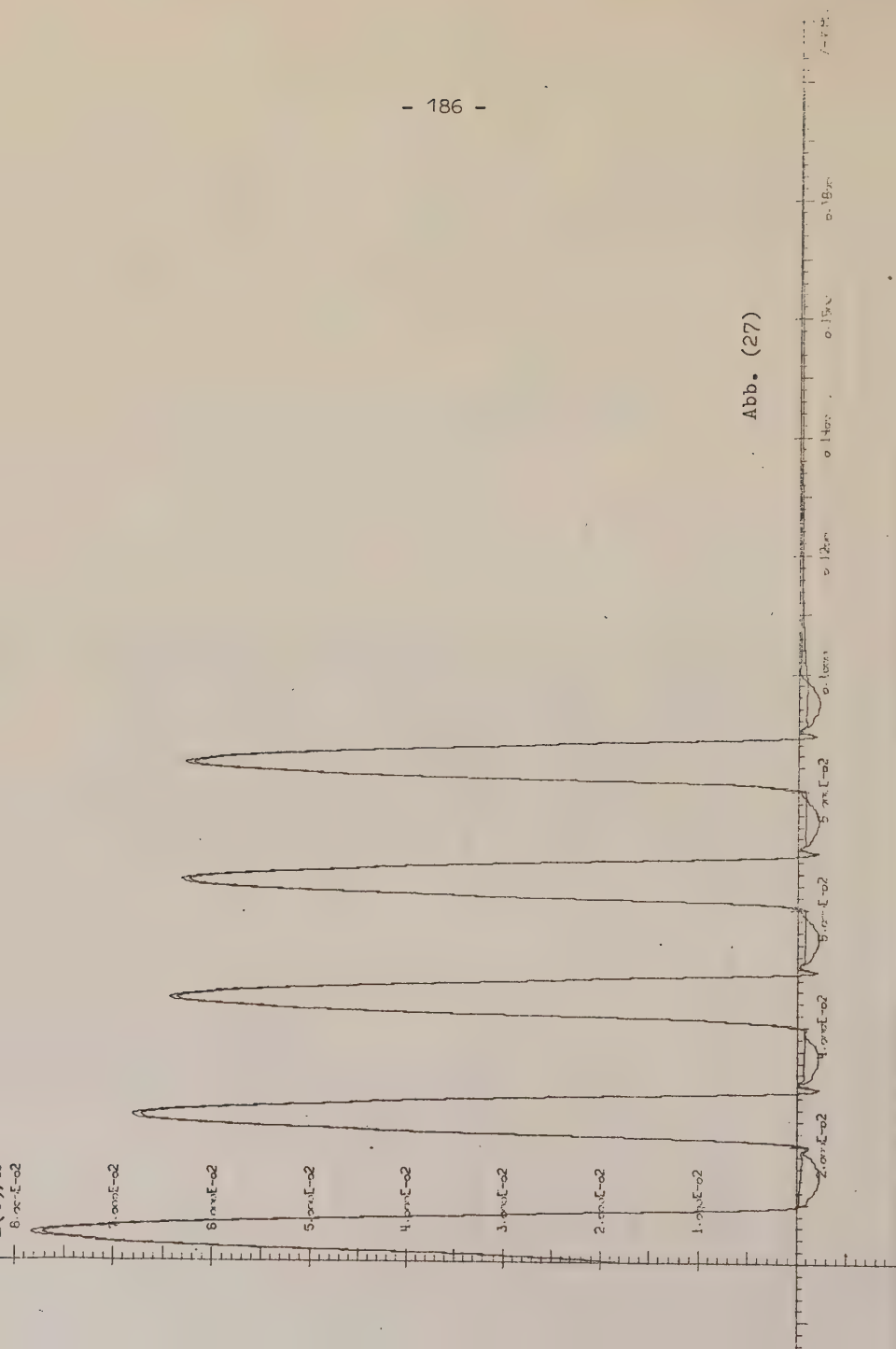
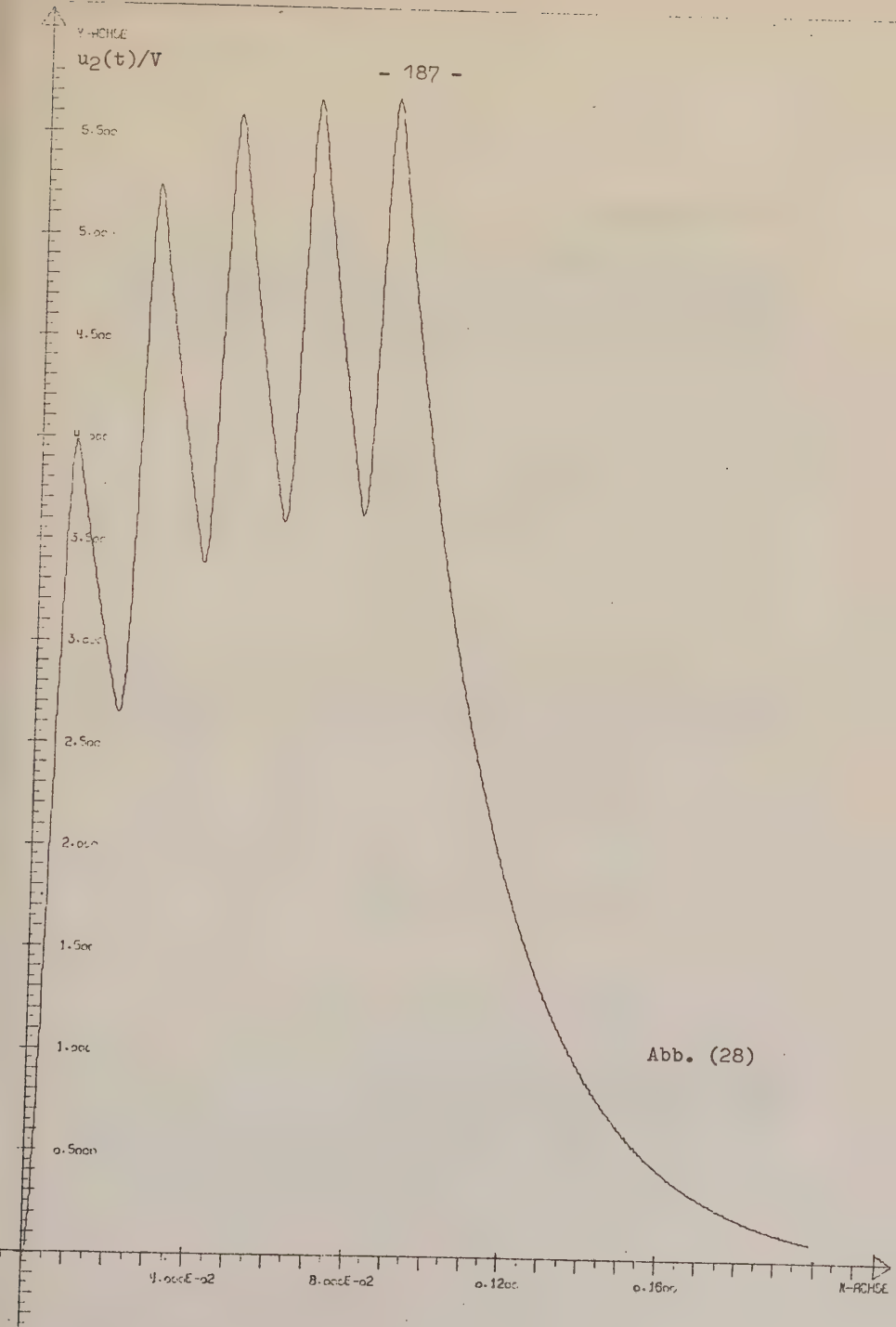


Abb. (27)



3.1.4 Beispiel K4

Gegeben sei das Rückkopplungssystem der Abb. (29) mit der skalaren Erregung e_0 gemäß Abb. (31) und der skalaren Systemreaktion r . Die Systemfunktionen e und f seien Vektorfunktionen der Dimension zwei (d.h. $n = 2$).

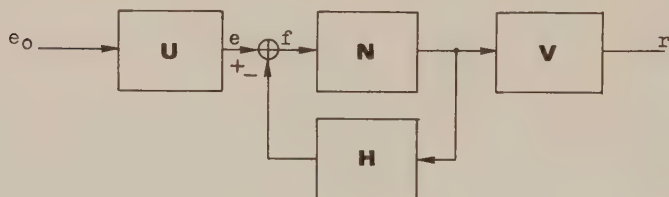


Abb. (29)

Die linearen Operatoren **U**, **H** und **V** seien durch die Matrizen von Übertragungsfunktionen

$$U(p) = \begin{pmatrix} U_{11}(p) \\ U_{21}(p) \end{pmatrix}$$

$$H(p) = \begin{pmatrix} H_{11}(p) & H_{12}(p) \\ H_{21}(p) & H_{22}(p) \end{pmatrix}$$

und

$$V(p) = (V_1(p), V_2(p))$$

definiert, wobei die skalaren Übertragungsfunktionen folgendermaßen gegeben sind:

$$U_1(p) = U_2(p) = \frac{1}{p+0.5}$$

$$H_{11}(p) = H_{22}(p) = (p^5 + 7.2p^4 + 16.4p^3 + 17.8p^2 + 15.6p + 2.1)/D(p)$$

$$H_{12}(p) = H_{21}(p) = -(p^2 + 0.3p + 3.)/D(p)$$

$$D(p) = p^5 + 2.5p^4 + 5p^3 + 5.5p^2 + 3p + 2$$

$$V_1(p) = V_2(p) = \frac{1}{p^2 + 0.2p + 0.5}.$$

Der nichtlineare Operator **N** sei durch die Vektorfunktion

$$\psi(x_1, x_2) = \begin{pmatrix} \psi_1(x_1) \\ \psi_2(x_2) \end{pmatrix}$$

gemäß Abb. (30 a,b) gegeben (der Übergang in den Sättigungsbereich wird dabei durch Parabeläste beschrieben), und es ist $\alpha_1 = \alpha_2 = 0$, $\beta_1 = 1$ und $\beta_2 = 0.5$.

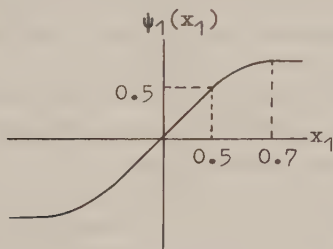


Abb. (30a)

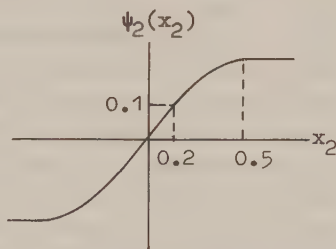


Abb. (30b)

Es ist $r_0 = 0.9$, und damit kann unser Verfahren angewandt werden. In den Abb. (32 a,b) sind die Komponenten der am Rückkopplungssystem wirkenden Erregung e dargestellt. Die Abb. (33 a,b) zeigen die Komponenten der ersten Näherung $\hat{f}_1$ von $\hat{f}$

und die Abb. (34 a,b) die Komponenten der letzten beiden Iterationsschritte. Die Systemreaktion r ist schließlich aus Abb. (35) ersichtlich. Da die letzten beiden Näherungsfunktionen visuell nicht voneinander unterscheidbar sind, ist die Genauigkeit des Verfahrens gut. Die Konvergenzgeschwindigkeit entspricht in etwa der Kontraktionsschranke r_0 . (Es waren 20 Iterationsschritte notwendig.) In Abb. (36) ist dargestellt, wie die Kontraktionsschranken $r(\mathbf{C}, \mathbf{T}(\mathbf{C}))$ vom Transformationsoperator $\mathbf{C}$ abhängt (siehe Abschnitt II.2.4.3). Wegen $n=2$ ist hier $\mathbf{C}$ durch $c_1, c_2 \in \mathbf{R}$ festgelegt. Wie man sieht, ist die Wahl der Transformationsoperatoren $\mathbf{C}_0$ und $\mathbf{T}_0$ optimal.

Durch eine leichte Veränderung des Rückkopplungssystems erhält man interessante Aufschlüsse: Wählt man die Konstante im Zählerpolynom von $H_{11}(p)$ und $H_{22}(p)$ statt 2.1 zu 2.0 und setzt man $\beta_1 = \beta_2 = 2$, so wird $r_0 = 1$. Dieser Fall stellt also eine Grenze zwischen Anwendbarkeit und Nichtanwendbarkeit unserer Theorie dar. Nimmt man an, daß die nichtlinearen Kennlinien nicht in den Sättigungsbereich hinein angesteuert werden (das kann durch hinreichend "kleine" Erregungsfunktionen immer erreicht werden), so wirkt das System wie ein lineares System. Man kann leicht zeigen, daß die Übertragungsfunktion dieses linearen Systems bei $p=0$ einen einfachen Pol besitzt, und damit hat das System integrierendes Verhalten. Das bedeutet aber, daß die Systemreaktion nicht abklingen wird. Deshalb ist nicht zu erwarten, daß sie quadratisch integrierbar ist. Daß für diesen Grenzfall keine quadratisch integrierbare Systemreaktion existiert ist ein weiterer Beweis für die Qualität des Existenzkriteriums. Unser Verfahren wurde auch auf diesen Grenzfall angewandt, und wie zu erwarten, war die Konvergenz schlecht. In Abb. (37) sind die letzten beiden Iterationsschritte (nur eine Komponente der Vektorfunktion) dargestellt, und man erkennt deutlich, wie die Systemfunktion beim Austritt aus dem Sättigungsbereich nicht mehr abnimmt.

PHYSIKALISCHE ERREGUNG

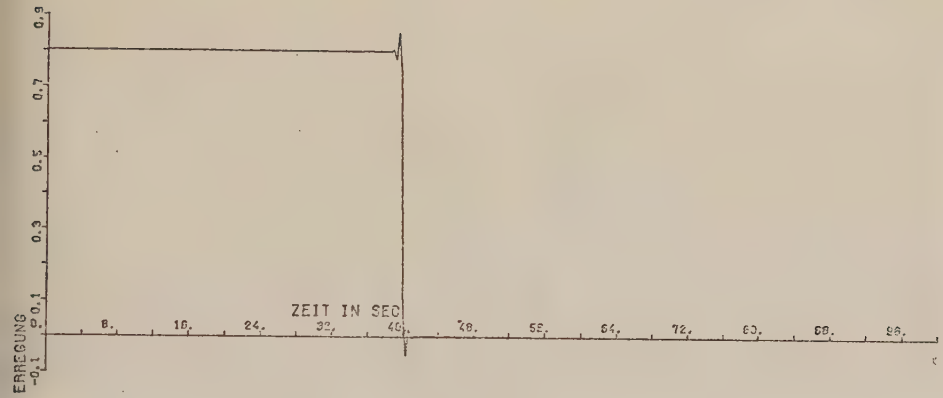


Abb. (31)

ERREGUNG E1 AM RUECKKOPPLUNGSSYSTEM.

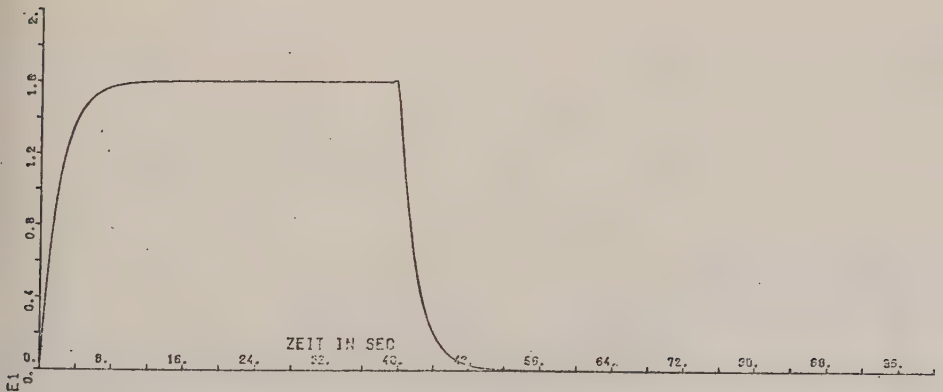


Abb. (32a)

ERREGUNG E2 AM RUECKKOPPLUNGSSYSTEM

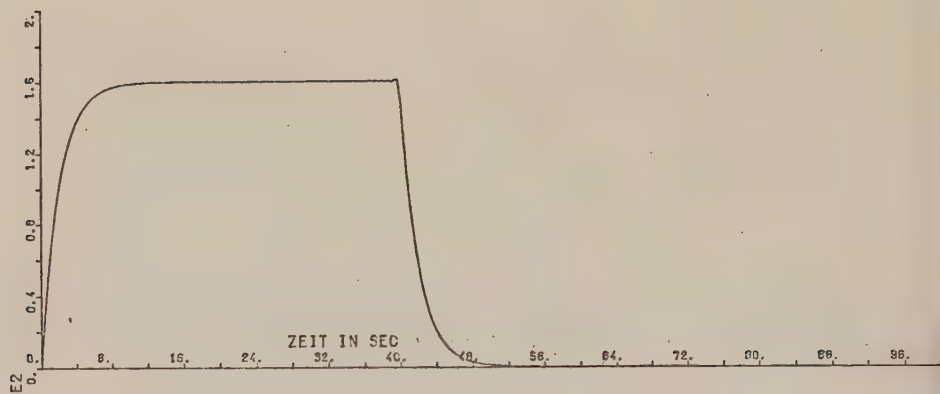


Abb. (32b)

ERREGUNG E1S AM TRANSFORMIERTEN R-SYSTEM

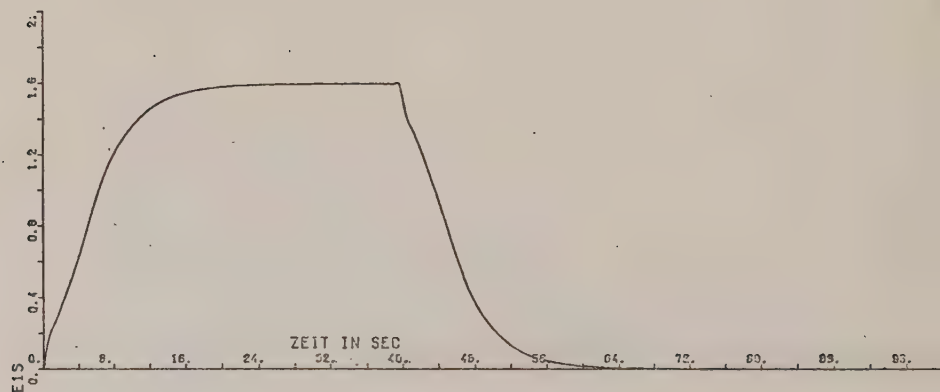


Abb. (33a)

ERREGUNG E2S AM TRANSFORMIERTEN R-SYSTEM

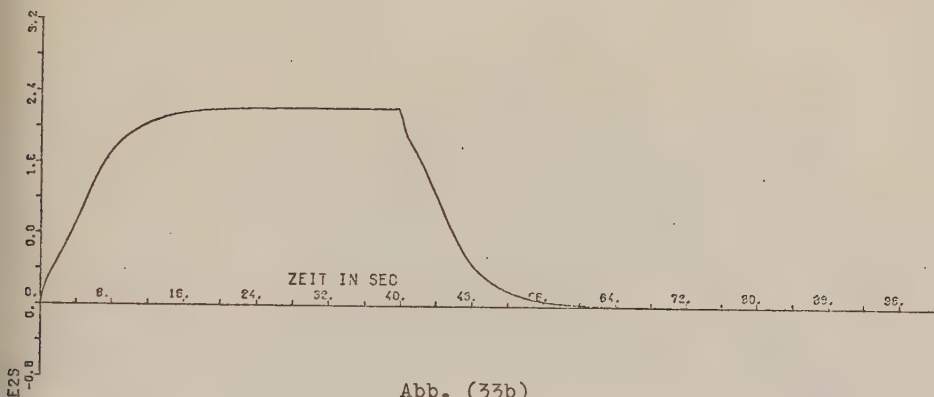


Abb. (33b)

LETZTEN BEIDEN ITERATIONEN

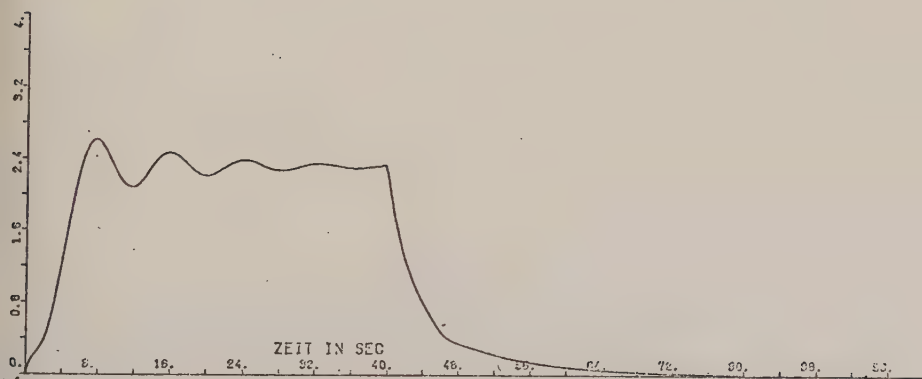


Abb. (34a)

LETZTEN BEIDEN ITERATIONEN

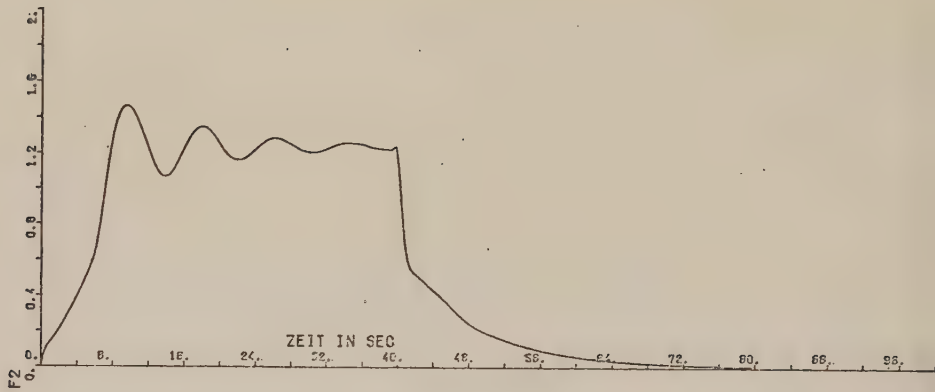


Abb. (34b)

ANTWORT DES NICHTLINEAREN SYSTEMS

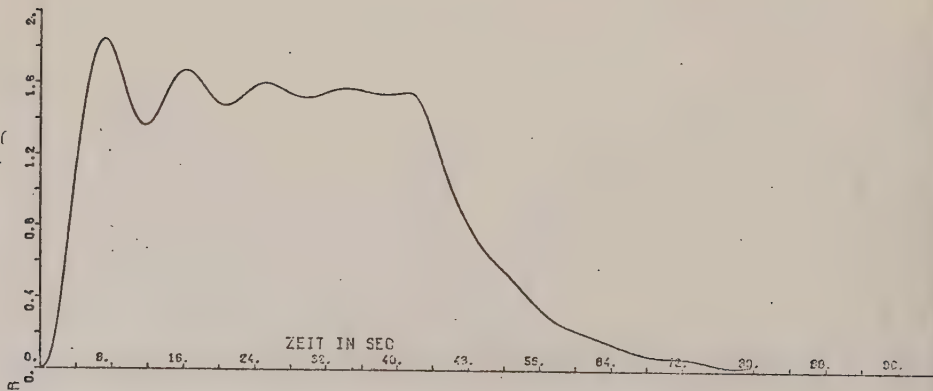


Abb. (35)

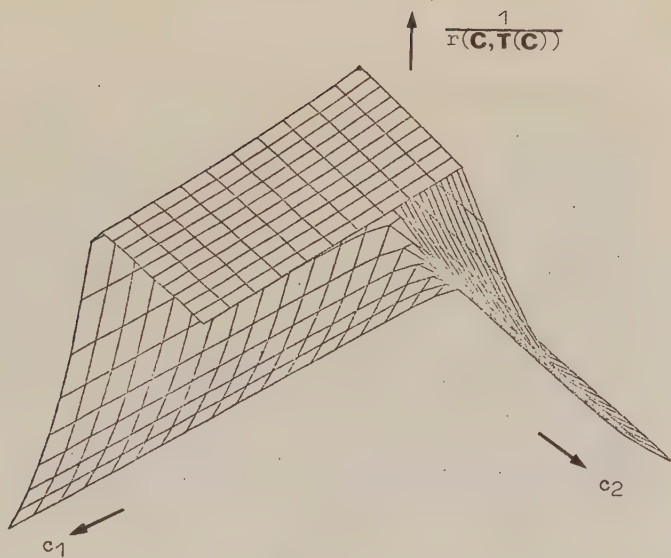


Abb. (36)

LETZTEN BEIDEN ITERATIONEN

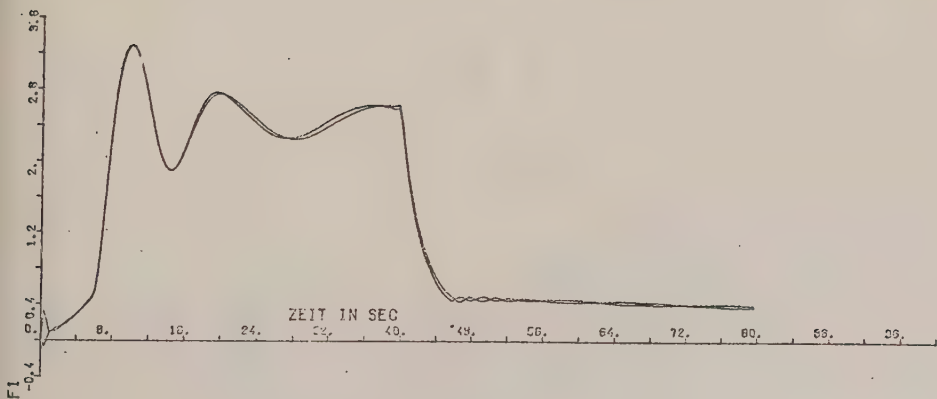


Abb. (37)

3.1.5 Beispiel K5

Gegeben sei eine aktive RC-Schaltung gemäß Abb. (40), bei der die mittleren beiden Verstärker, wie in Abb. (39 a,b) gezeigt, nichtlinear seien. Diese Schaltung werde durch die Spannung u_1 gemäß Abb. (41) erregt, und u_2 ist die gesuchte Systemfunktion. Nach Abschnitt III.3.2 kann dieses Netzwerk durch das äquivalente Systemblockschaltbild der Abb. (38) dargestellt werden. Die linearen Operatoren können durch Übertragungsfunktionen repräsentiert werden, welche anhand einer einfachen Analyse ermittelt werden können, und der nichtlineare Operator **N** ist gemäß Abb. (39) definiert, wobei $\alpha_1 = \alpha_2 = 0$, $\beta_1 = 1$ und $\beta_2 = 2$ gilt (der Übergang in den Sättigungsbereich sei durch Parabeläste gegeben).

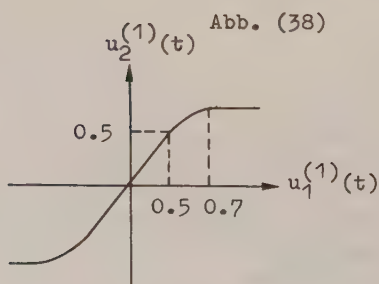
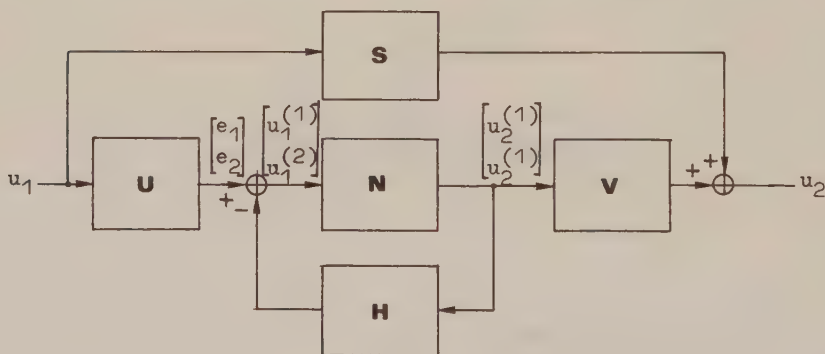


Abb. (39a)

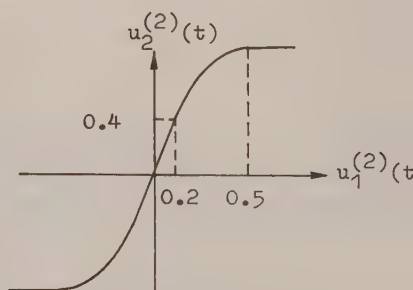


Abb. (39b)

Wegen $r_0 = 0.84$ ist unser Verfahren anwendbar. Die Abb. (42 a) und (42 b) zeigen den Verlauf der Erregung am Rückkopplungssystem. Die Abb. (43 a) und (43 b) stellen die ersten Approximationsfunktionen, die Abb. (44 a) und (44 b) die letzten beiden Approximationsfunktionen der Eingangsspannungen an den beiden nichtlinearen Verstärkern und die Abb. (44 a) und (44 b) die Ausgangsspannungen an den beiden nichtlinearen Verstärkern dar. Schließlich zeigt Abb. (46) die Ausgangsspannung u_2 des Netzwerkes. (Wären alle Verstärker linear, so würde das Netzwerk als Impulsantwort eine angenäherte Dreiecksfunktion liefern.)

Auch hier entspricht die Konvergenzgeschwindigkeit in etwa der Kontraktionsschranke r_0 , und die Genauigkeit ist, wie aus den Abb. (44 a) und (44 b) ersichtlich, sehr groß.

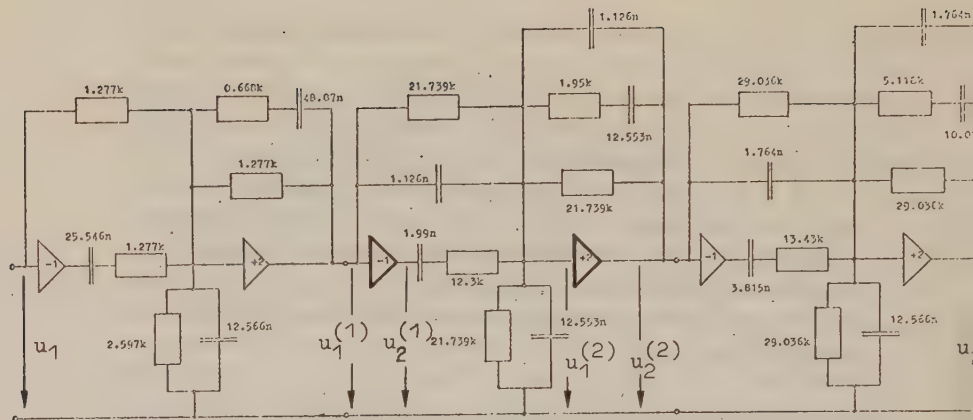


Abb. (40)

PHYSIKALISCHE ERREGUNG

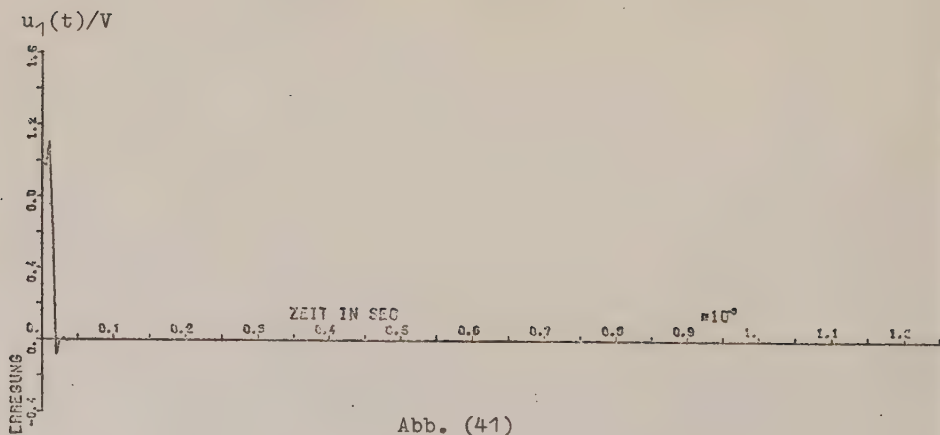


Abb. (41)

ERREGUNG E1 AN RUECKKOPPLUNGSSYSTEM (in V)

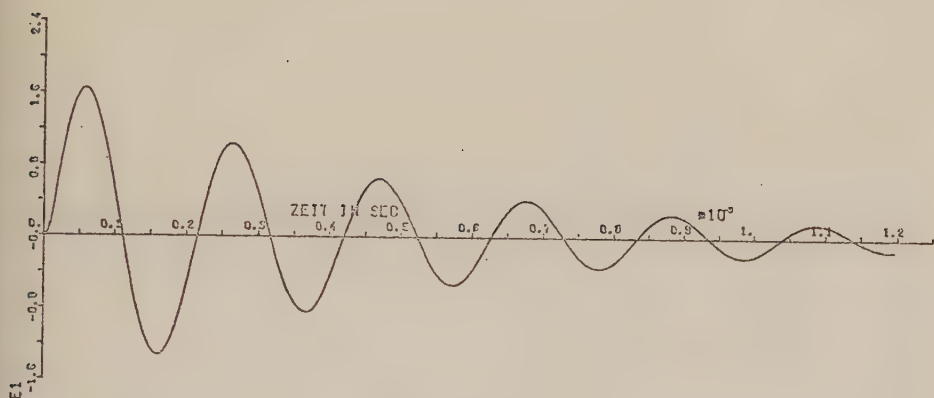


Abb. (42a)

ERREGUNG E2 AM RUECKKOPPLUNGSSYSTEM (in V)

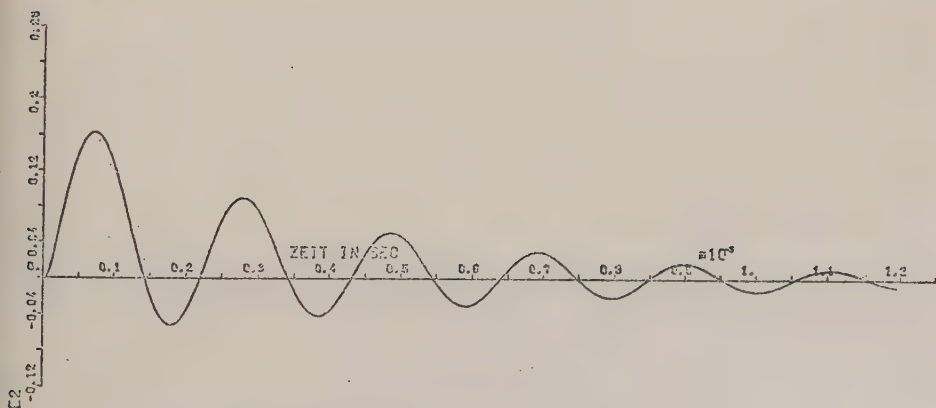


Abb. (42b)

ERREGUNG E1S AM TRANSFORMIERTEN R-SYSTEM (in V)

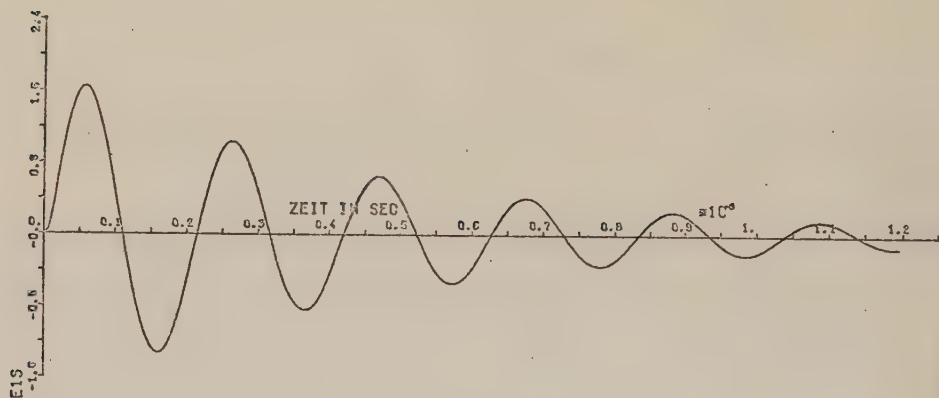


Abb. (43a)

ERREGUNG E2S AM TRANSFORMIERTEN R-SYSTEM (in V)

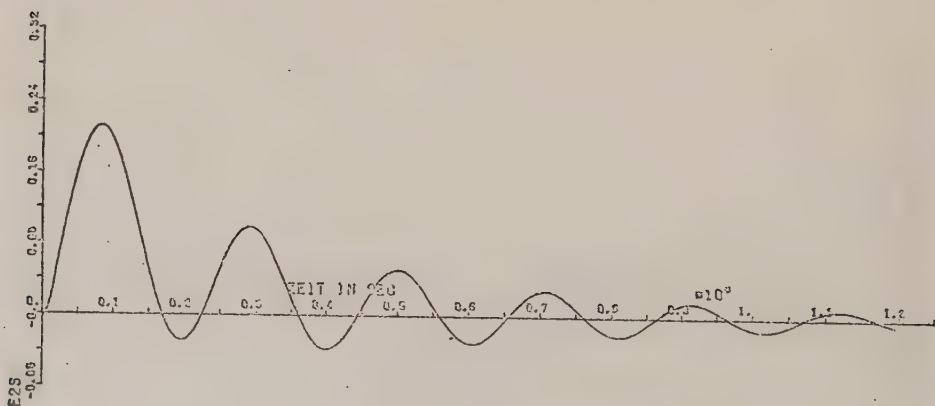


Abb. (43b)

$$u_1^{(1)}(t)/V$$

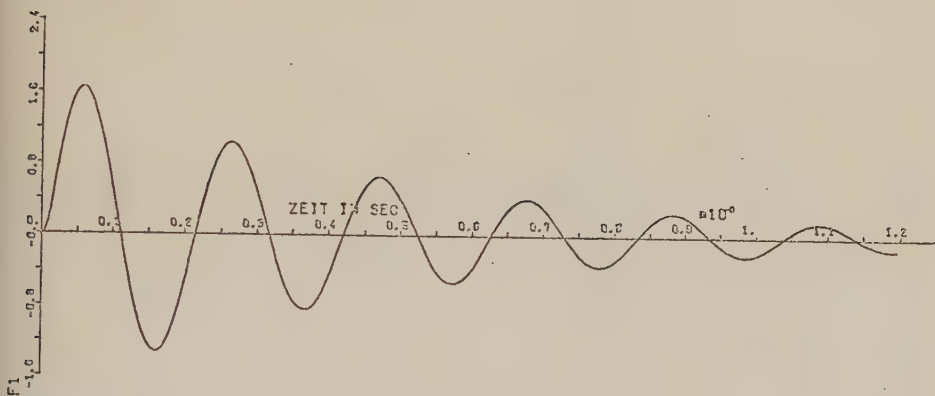


Abb. (44a)

LETZTEN BEIDEN ITERATIONEN

$$u_1^{(2)}(t)/V$$

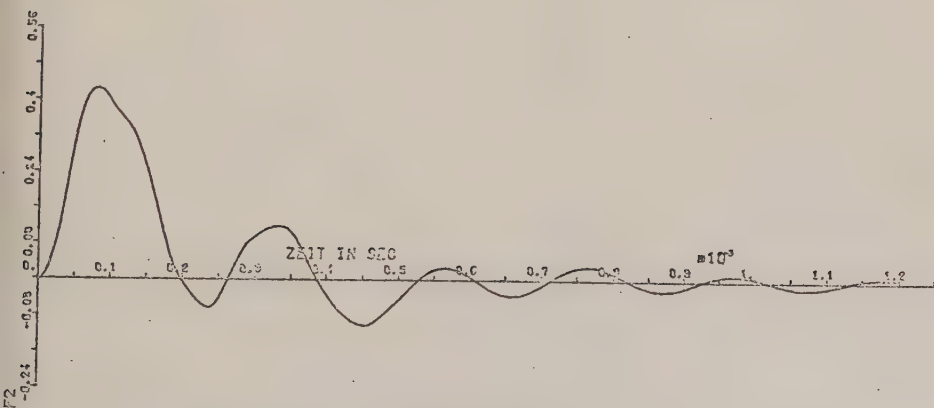


Abb. (44b)

$$u_2^{(1)}(t)/V$$

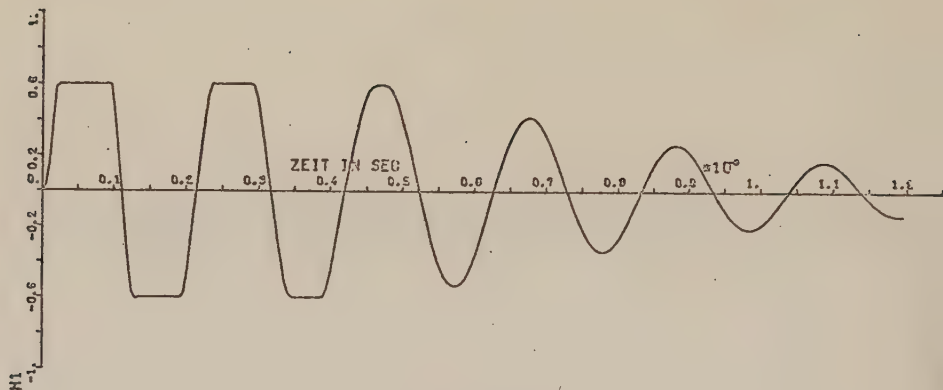


Abb. (45a)

$$u_2^{(2)}(t)/V$$

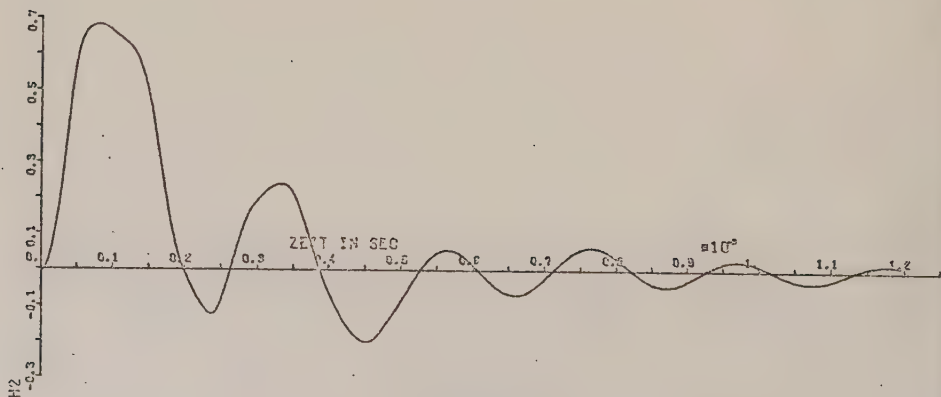


Abb. (45b)

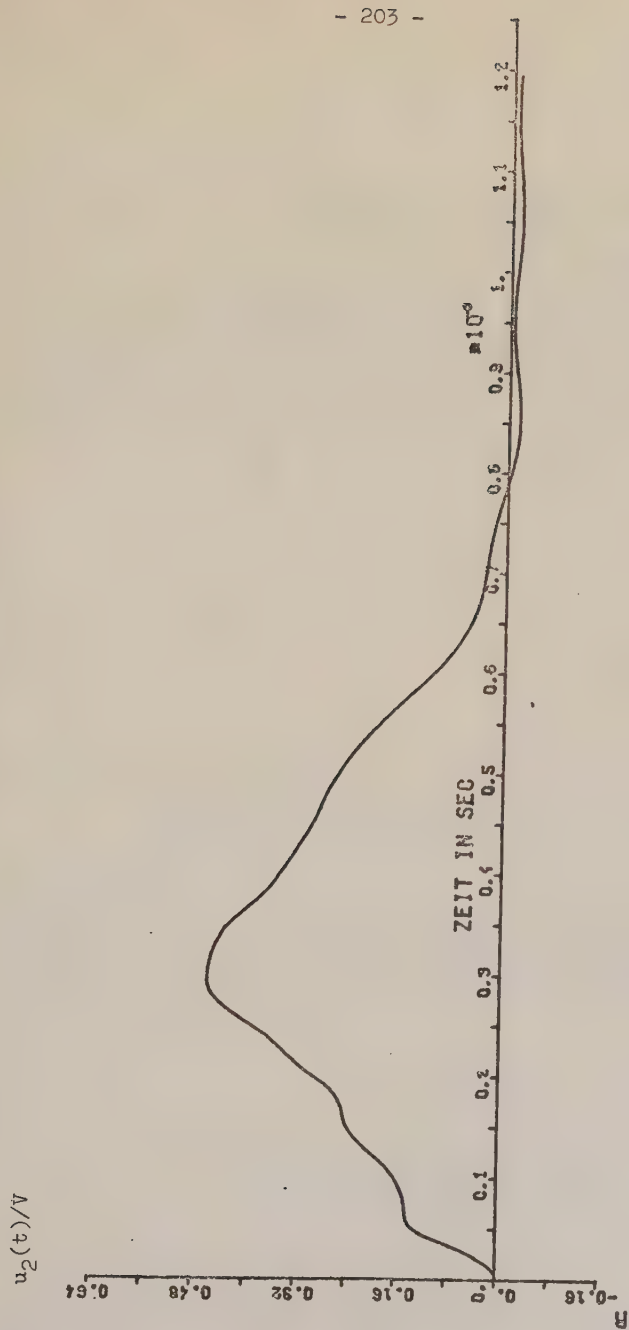


Abb. (46)

3.2 Systeme mit verteilten Parametern

3.2.1 Beispiel V1

Gegeben sei das Rückkopplungssystem der Abb. (47) mit der skalaren Erregung e gemäß Abb. (48) und der skalaren Systemreaktion f .

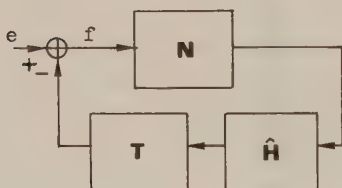


Abb. (47)

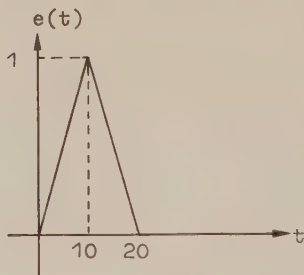


Abb. (48)

Der lineare Operator $\hat{H}$ sei durch die Übertragungsfunktion

$$\hat{H}(p) = \frac{p^2 + 3.8p + 2.4}{p^5 + 7.2p^4 + 16.4p^3 + 17.8p^2 + 15.6p + 2.}$$

definiert, und der lineare Operator T sei ein Verzögerungsglied mit der Verzögerungszeit $T_0 = 5$. Der nichtlineare Operator N sei gemäß Abb. (49) mit $\alpha = 0$ und $\beta = 1.5$ definiert. (Der Übergang vom Sättigungsbereich sei durch Parabeläste gegeben.)

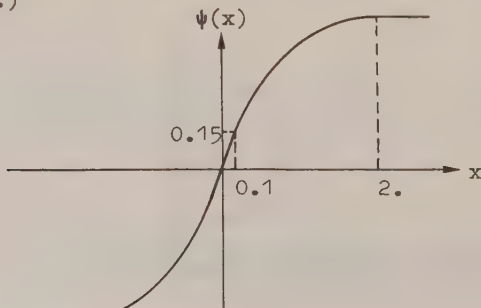


Abb. (49)

Wie man aus dem Frequenzgang der Abb. (50) erkennt, ist unser Verfahren anwendbar, und es gilt $r_0 = 0.67$. Abb. (51) zeigt die erste Approximationsfunktion und Abb. (52) die folgenden Approximationsfunktionen für die gesuchte Systemreaktion f . In Abb. (53) sind schließlich die letzten beiden Approximationsfunktionen dargestellt. Man erkennt auch hier die große Genauigkeit des Analyseverfahrens. Wieder entspricht die Konvergenzgeschwindigkeit in etwa der Kontraktionsschranke.

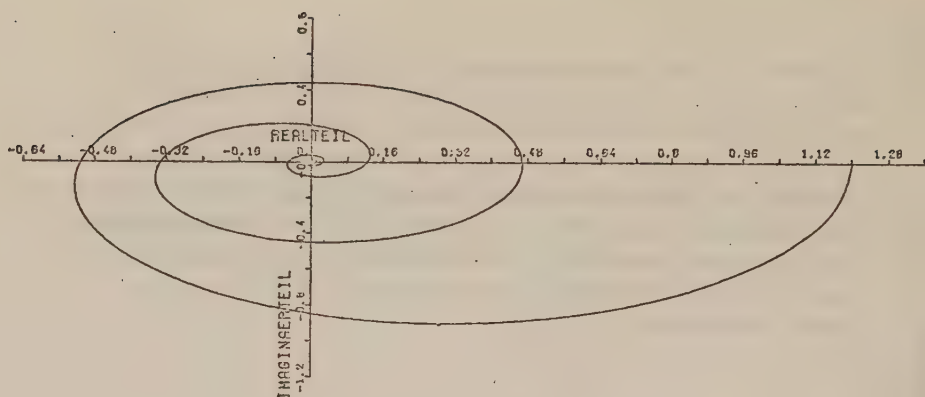


Abb. (50)

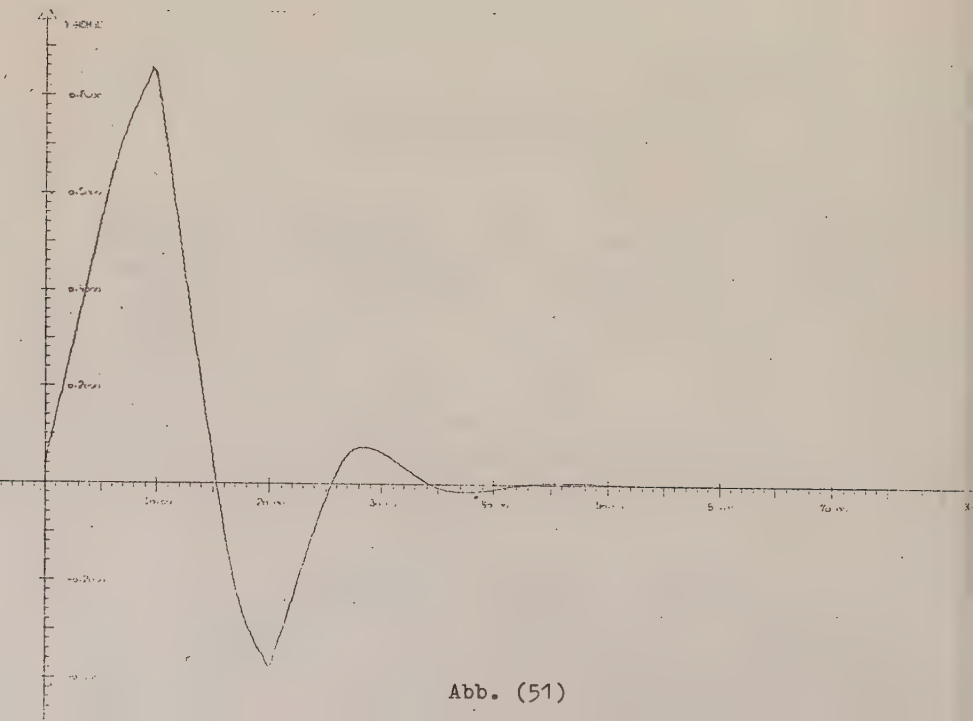


Abb. (51)

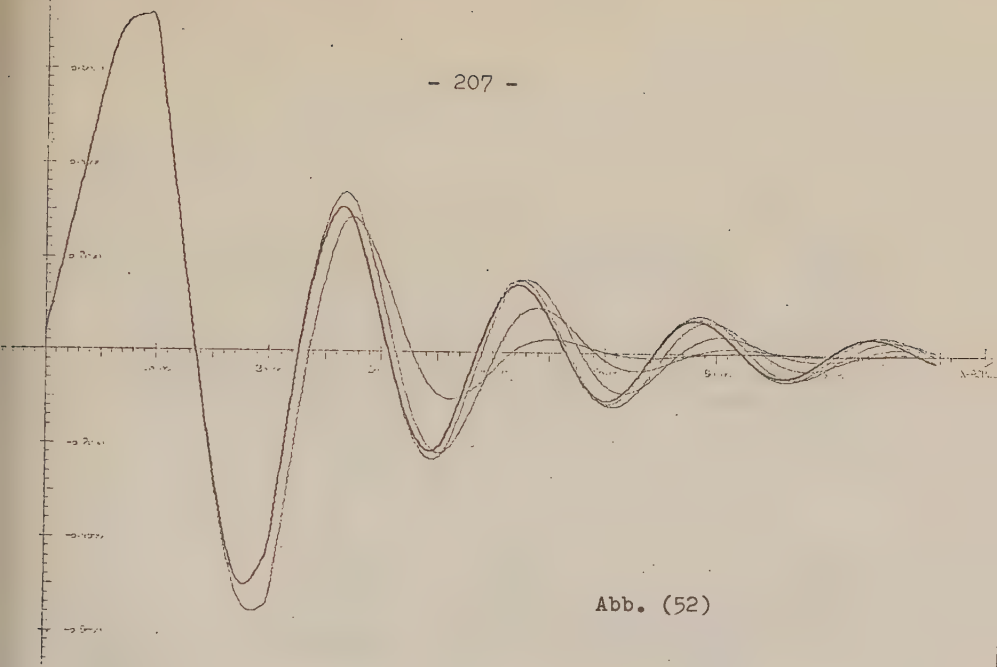


Abb. (52)

ÜBERSTICH VON BEISPIEL NR. V1

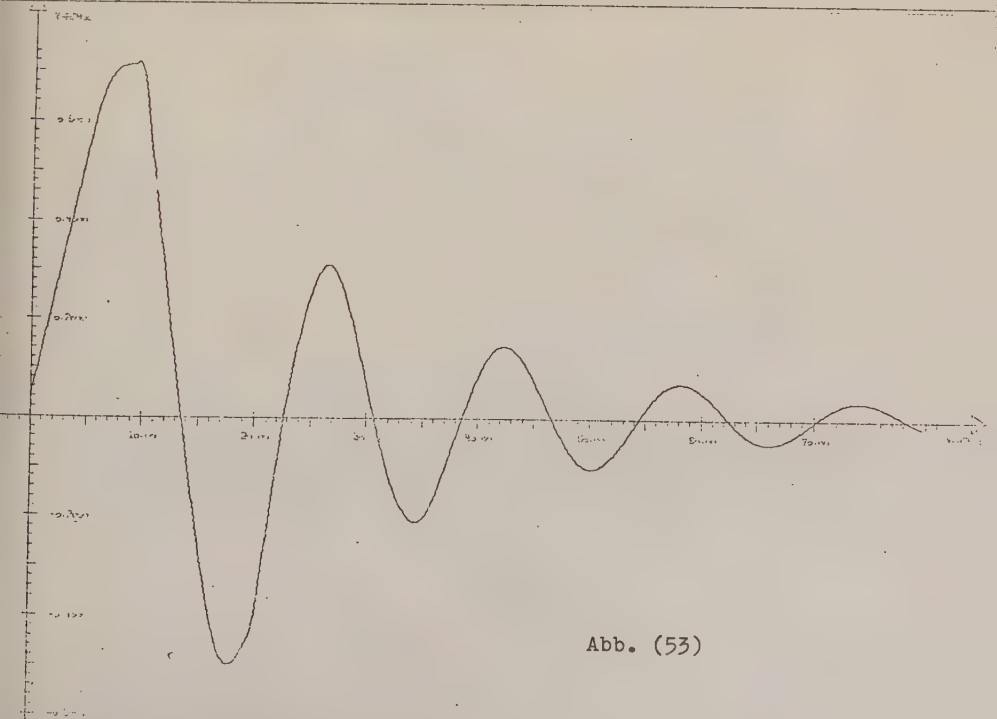


Abb. (53)

3.2.2 Beispiel V2

Gegeben sei das Netzwerk der Abb. (54), das aus einer homogenen elektrischen Leitung mit den Leitungsbelägen $L' = 0.1 \text{ H/km}$, $C' = 0.4 \text{ } \mu\text{F/km}$, $R' = 50 \text{ } \Omega/\text{km}$ und $G' = 10 \text{ nS/km}$ besteht, die durch eine Spannungsquelle u_0 gemäß Abb. (55) erregt wird und mit dem durch Abb. (56) definierten nicht-linearen Widerstand abgeschlossen ist. (Dabei ist $a = 0.5$ und $\beta = 10^{-1} \text{ S}$.)

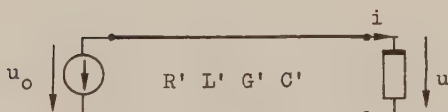


Abb. (54)

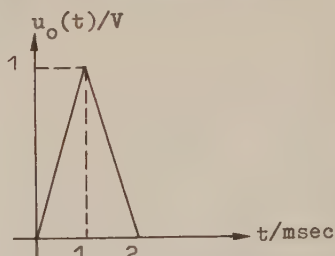


Abb. (55)

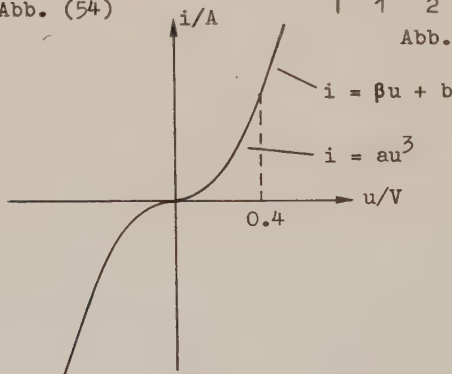


Abb. (56)

Dieses Netzwerk kann gemäß Abschnitt III.3.1 durch das Systemblockschaltbild der Abb. (57) dargestellt werden, wobei die linearen Operatoren $\mathbf{U}$ und $\mathbf{H}$ durch die Übertragungsfunktionen

$$U(p) = \frac{1}{\cosh(\gamma l)}$$

und

$$H(p) = Z_W \cdot \tanh(\gamma l)$$

definiert sind.

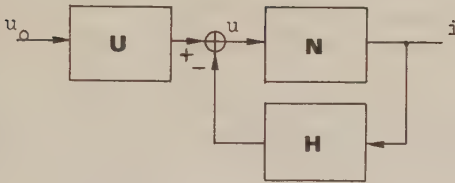


Abb. (57)

Dabei sind das Leitungsmaß γ und der Wellenwiderstand Z_W durch

$$\gamma = \sqrt{(R' + pL')(G' + pC')}$$

und

$$Z_W = \sqrt{\frac{R' + pL'}{G' + pC'}}$$

gegeben, und die Leitungslänge l beträgt 10 km. Genau genommen ist unsere Theorie hier nicht anwendbar, da die der Übertragungsfunktion $H(p)$ zugeordnete Impulsantwort i.a. einen Impulskamm enthält. Mittels eines subtileren Beweisverfahrens läßt sich aber zeigen, daß unter schwachen Bedingungen unsere Theorie auch für Impulsantworten dieser Art anwendbar ist. Wegen der Passivität von **H** und **N** kann unser Verfahren angewandt werden, und es ist $r_0 = 0.87$. Abb. (58) zeigt die erste Approximationsfunktion der gesuchten Ausgangsspannung u und Abb. (59) stellt die folgenden Approximationsfunktionen dar. In Abb. (60) sind die letzten beiden Approximationsfunktionen abgebildet.

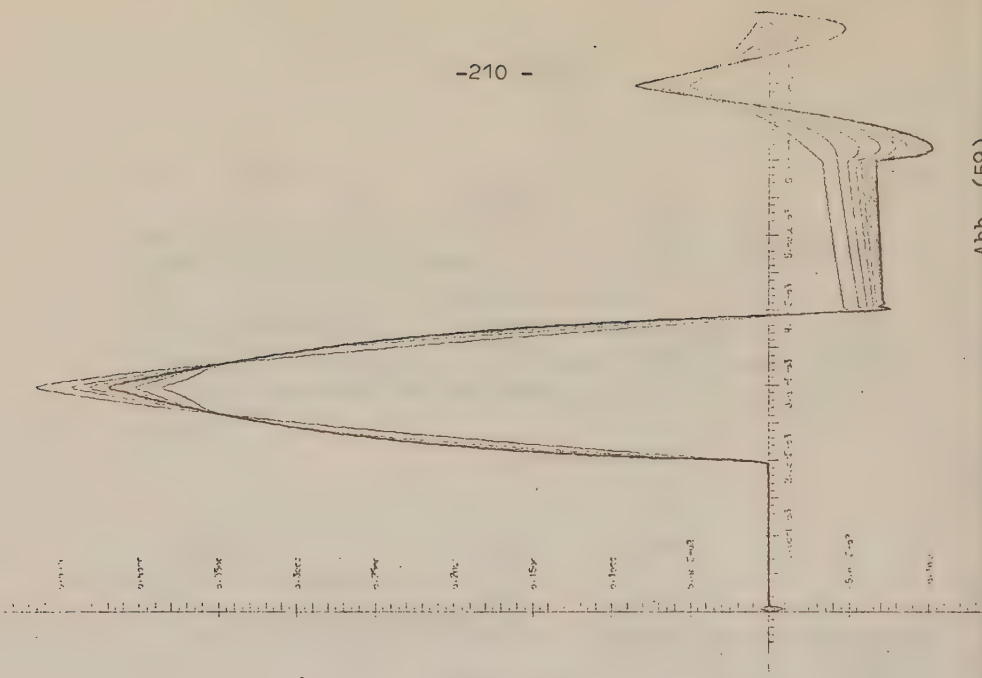


Abb. (58)

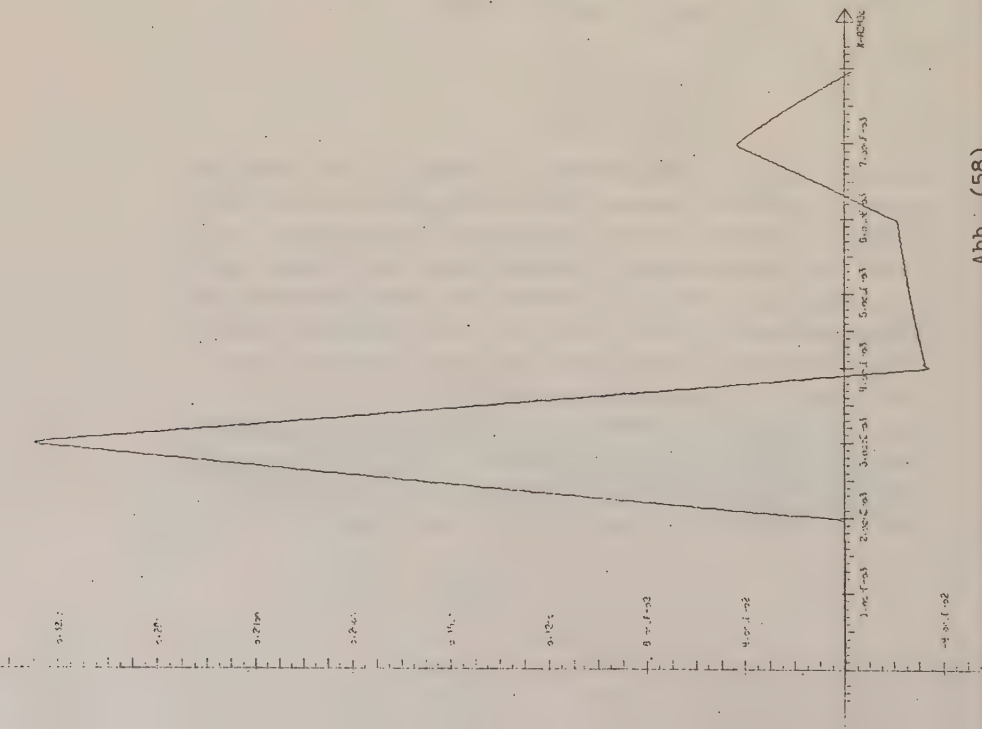


Abb. (59)

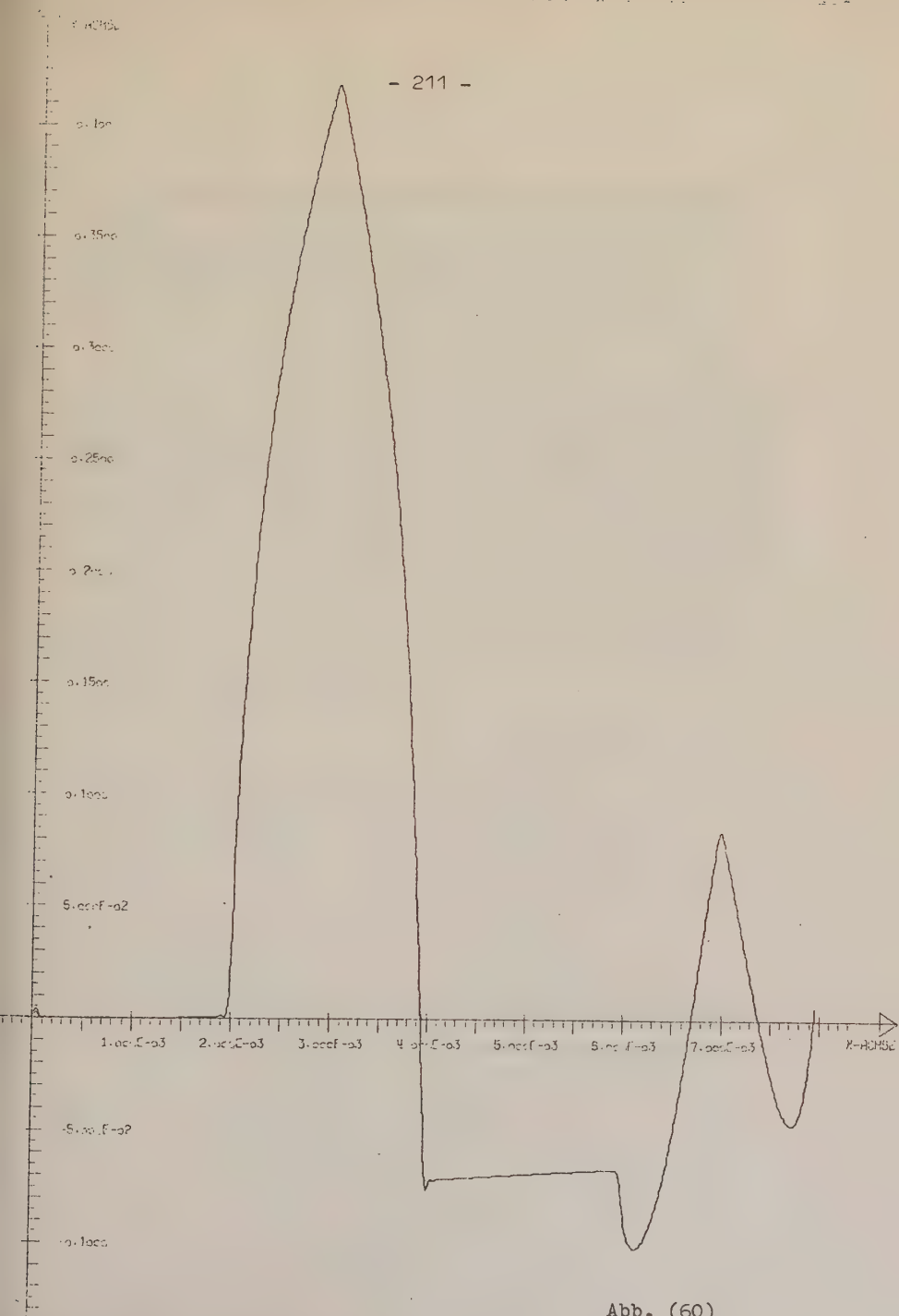


Abb. (60)

3.3 Vergleich mit dem Verfahren von Bulirsch-Stoer

In diesem Abschnitt vergleichen wir unser Analyseverfahren mit dem bekannten und leistungsfähigen Integrationsverfahren von Bulirsch-Stoer [1], wobei das zu analysierende System gemäß Abb. (61) gegeben sei. Sowohl die Systemreaktion f als auch die Systemerregung e seien dabei skalare Funktionen. Ferner setzen wir voraus, daß die den linearen Operator H beschreibende Übertragungsfunktion $H(p)$ eine gebrochen rationale Funktion ist (nur dann kann eine gewöhnliche Differentialgleichung für das System formuliert werden), und daß diese in der abgeschlossenen rechten p -Halbebene einschließlich des Punktes $p = \infty$ polfrei ist. Der nichtlineare Operator N sei durch die nichtlineare Funktion $\psi(u)$ gemäß Abb. (62) definiert, und dabei ist $\alpha = 0$ und $\beta = 0.9$. (Der Übergang in den Sättigungsbereich sei durch Parabeläste gegeben.)

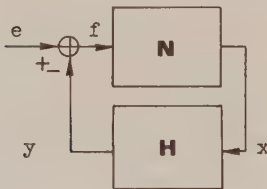


Abb. (61)

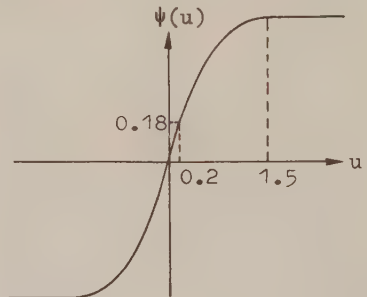


Abb. (62)

Um das Verfahren von Bulirsch-Stoer anwenden zu können, müssen wir eine Zustandsraumgleichung der Form

$$\begin{aligned}\dot{z} &= f_1(z, e) \\ f &= f_2(z, e)\end{aligned}\tag{6}$$

aufstellen, die das System beschreibt. Zunächst können wir aus $H(p)$ eine den linearen Operator $\mathbf{H}$ definierende Zustandsdarstellung ableiten [22], wobei x das Eingangssignal, y das Ausgangssignal und z der Zustandsvektor ist:

$$\dot{z} = Az + Bx$$

$$y = Cz + dx.$$

Da es im Fall $d \neq 0$ nicht direkt möglich ist, eine Zustandsgleichung der Gl. (6) zu formulieren, versuchen wir gemäß Abb. (63 a), (63 b) und (63 c) den Einfluß von d in den nichtlinearen Operator zu integrieren. Der verbleibende lineare Restoperator $\tilde{\mathbf{H}}$ ist dann durch die Zustandsgleichungen

$$\dot{z} = Az + Bx$$

$$y = Cz$$

definiert, und jetzt hängt die Ausgangsgröße y nur noch über z von der Eingangsgröße x ab.

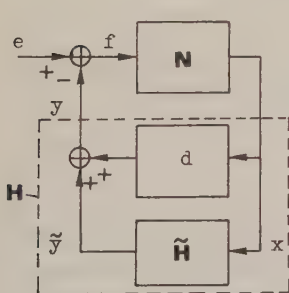


Abb. (63a)

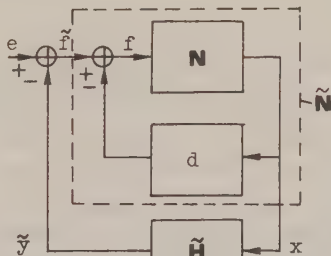


Abb. (63b)

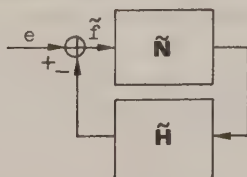


Abb. (63c)

Der Operator $\tilde{\mathbf{N}} = \mathbf{N}(\mathbf{I} + d\mathbf{N})^{-1}$ existiert sicher, wenn die Bedingungen für die Anwendbarkeit unseres Analyseverfahrens erfüllt sind, da $\mathbf{N}$ den Systemoperator $\mathbf{N}(\mathbf{I} + \mathbf{H}\mathbf{N})^{-1}$ für "hochfrequente" Erregungen darstellt. Die gesuchte Zustandsraumgleichung lautet jetzt

$$\begin{aligned}\dot{\mathbf{z}} &= \mathbf{A}\mathbf{z} + \mathbf{B}\tilde{\Psi}(\mathbf{e} - \mathbf{C}\mathbf{z}) \\ \mathbf{f} &= \mathbf{e} - \mathbf{C}\mathbf{z} - d\tilde{\Psi}(\mathbf{e} - \mathbf{C}\mathbf{z}),\end{aligned}$$

wobei $\tilde{\mathbf{N}}$ durch die Funktion $\tilde{\Psi}(u)$ gegeben ist. Die Funktion $\tilde{\Psi}(u)$ ergibt sich als Hintereinanderschaltung von Funktionen, d.h.

$$\tilde{\Psi}(u) = \psi(\phi(u)),$$

wobei $\mathbf{w} = \phi(u)$ für jedes vorgegebene u durch die Lösung der Gleichung

$$\mathbf{w} + d\tilde{\Psi}(\mathbf{w}) = u$$

ermittelt werden kann (z.B. mit unserem Verfahren im Sinne von Abschnitt III.3.4).

Beim Vergleich unseres Verfahrens (kurz ITER genannt) mit dem Verfahren von Bulirsch-Stoer (kurz BULSTO genannt) gehen wir von Allpaß-Übertragungsfunktionen der Form

$$H(p) = \left(\frac{p-1}{p+1} \right)^m$$

mit beliebigem ganzzahligen $m \geq 2$ aus. (Durch diese Übertragungsfunktionen können Totzeitglieder mit der Totzeit $T_0 = 2m$ näherungsweise dargestellt werden.) Da der Frequenzgang $H(j\omega)$ ein Kreis mit Radius 1 um den Ursprung ist, folgt aus dem Frequenzgangkriterium die Anwendbarkeit unseres Verfahrens.

Beide Verfahren wurden für $m = 5, 10$ und 20 bezüglich Rechenzeit, Speicherbedarf (Rechenanlage TR 440) und Genauigkeit

verglichen, wobei die Systemerregung gemäß Abb. (65) gewählt wurde. Der Kernspeicherbedarf beträgt bei ITER 32 k und bei BULSTO 29 k. Die Rechenzeit ist aus der Tabelle der Abb. (64) zu entnehmen.

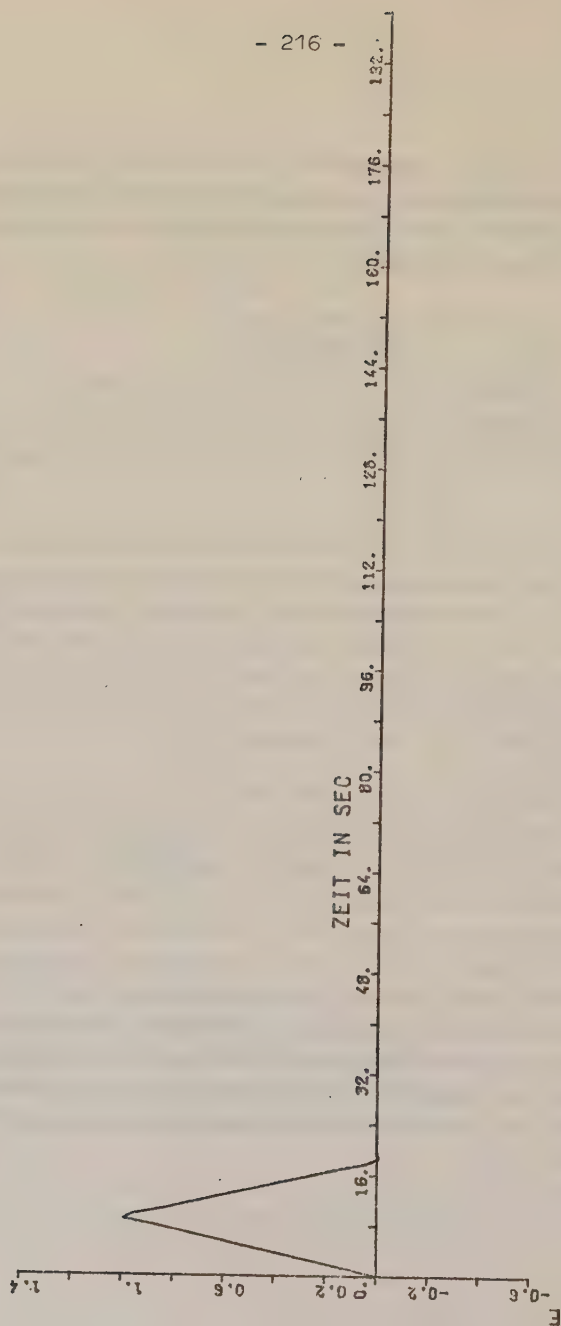
m	5	10	20	Rechenzeit in Minuten
BULSTO	0.25	4	8	
ITER	1	1	1	

Abb. (64)

Im Falle $m = 5$ und $m = 10$ liefern beide Verfahren Lösungsfunktionen, die visuell nicht unterscheidbar sind, was bedeutet, daß die Genauigkeit beider Verfahren sehr groß ist. Die Abb. (66) und (67) zeigen die Lösungsfunktionen beider Verfahren für $m = 10$, während die Abb. (68) und (69) die Lösungsfunktionen für $m = 20$ zeigen. Für $m = 20$ versagt das Verfahren BULSTO, und unser Verfahren liefert die zu erwartende Lösungsfunktion.

Der große Vorteil unseres Verfahrens liegt darin begründet, daß die Genauigkeit (bis auf den systematischen Fehler) durch den Abstand der letzten beiden Approximationsfunktionen abgeschätzt werden kann. Ferner können mit unserem Verfahren auch Systeme mit verteilten Parametern analysiert werden. Der Nachteil unseres Verfahrens besteht darin, daß es nur bei Erfülltsein bestimmter Bedingungen angewandt werden kann. Dieser Nachteil wird aber zum Teil dadurch wettgemacht, daß neben der Lösungsfunktion auch globale Eindeutigkeits-, Existenz- und Stetigkeitsaussagen geliefert werden.

PHYSIKALISCHE ERREGUNG



- 216 -

Abb. (65)

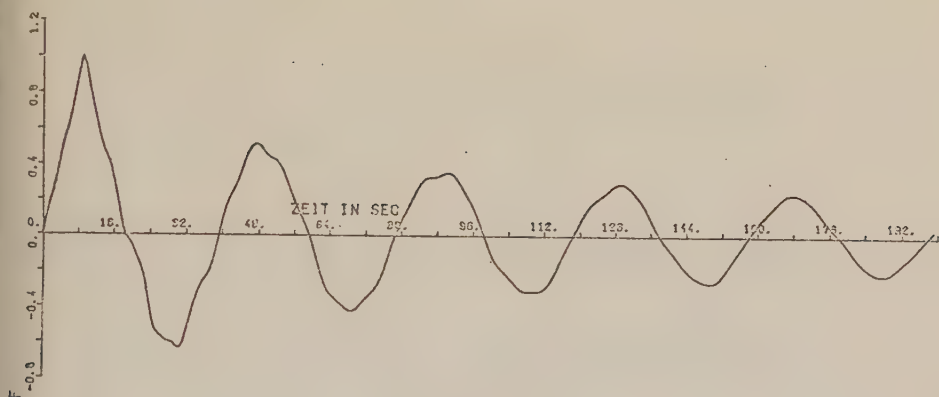


Abb. (66)

ANTWORT DES NICHTLINEAREN SYSTEMS

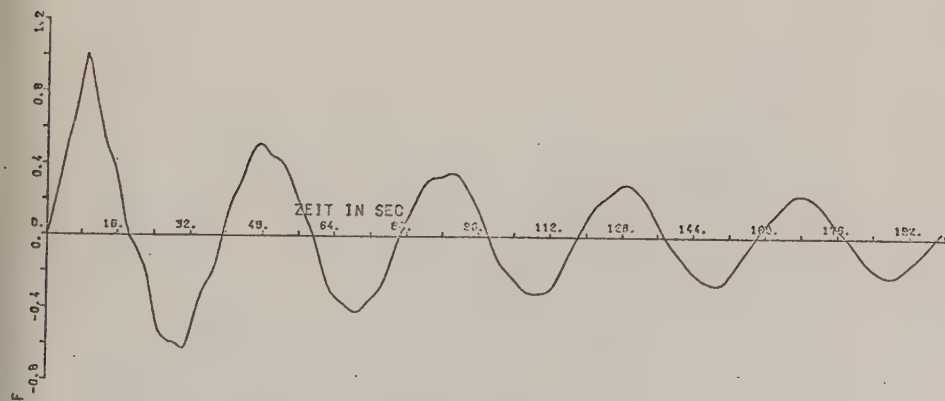


Abb. (67)

LETZTEN BEIDEN ITERATIONEN.

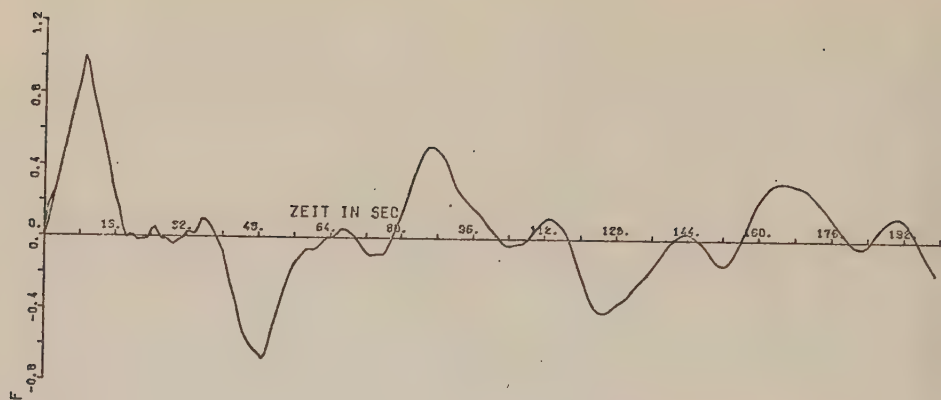


Abb. (68)

ANTICAT DES NICHTLINEAREN SYSTEMS

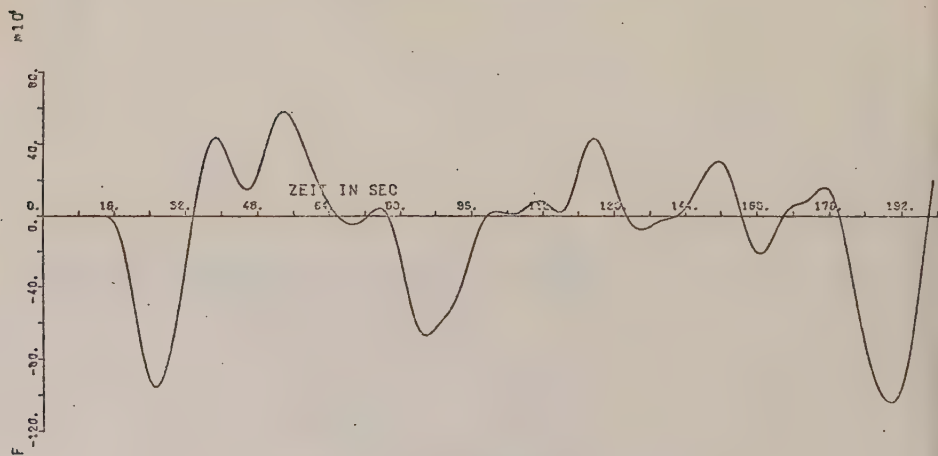


Abb. (69)

ZUSAMMENFASSUNG

Die vorliegende Arbeit beschäftigt sich mit der Analyse einer Klasse nichtlinearer Rückkopplungssysteme, die durch Operatorgleichungen der Art

$$e = f + \mathbf{H}\mathbf{N}f$$

und

$$e = \mathbf{N}f + \mathbf{H}f$$

beschrieben werden, wobei e die gegebene Systemerregung, f die gesuchte Systemreaktion, $\mathbf{H}$ einen linearen Operator und $\mathbf{N}$ einen nichtlinearen Operator darstellt. Die zentrale Aufgabe in dieser Arbeit ist es, Bedingungen anzugeben, unter denen eine eindeutige Lösung der Operatorgleichung existiert und diese auf einfache Weise berechnet werden kann. Diese Aufgabenstellung legt die Anwendung des Banachschen Fixpunktsatzes nahe, da dieser Existenzsatz konstruktiv ist und uns damit neben der Absicherung der Eindeutigkeit auch ein Iterationsverfahren zur Berechnung der Lösung liefert. Dabei werden die Operatorgleichungen in äquivalente Gleichungen transformiert, auf die dann unter den genannten Bedingungen der Fixpunktsatz angewandt werden kann. Ein wesentlicher Teil dieser Arbeit besteht aus der Untersuchung, auf welche Weise man diese Transformation am besten durchführt. Es wird eine Transformation vorgeschlagen, die am ehesten die Anwendbarkeit des Fixpunktsatzes und die schnellstmögliche Konvergenz des Iterationsverfahrens bezüglich einer sehr allgemeinen Klasse von Transformationen gewährleistet. Dabei wird im ersten Iterationsschritt die Lösung des in einem gewissen Sinne optimal linearisierten Gesamtsystems berechnet und danach erfolgen sukzessive Verbesserungen dieser Ausgangsnäherung. Diese Optimalitätsuntersuchungen erfolgen unter sehr allgemeinen Voraussetzungen, so daß die Analysetheorie auf die verschiedensten praktischen Probleme anwendbar ist. So werden die allgemeinen Aussagen technisch konkretisiert indem wir Übergangsvorgänge betrachten, wobei wir den linearen

Operator **H** als Faltungsoperator voraussetzen und den nicht-linearen Operator **N** durch einen Satz nichtlinearer Funktionen definieren. Wir erhalten damit Kriterien für die Anwendbarkeit unserer Theorie, die aus der Sicht des Ingenieurs leicht interpretiert werden können, da sie sich im wesentlichen auf die Fouriertransformierte der Impulsantwort des linearen Systems beziehen. Danach zeigen wir, daß diese spezialisierte Theorie auf die Analyse von elektrischen Netzwerken angewandt werden kann, die neben linearen Elementen nichtlineare zeitvariante Widerstände und gesteuerte Quellen enthält. Aufgrund unserer sehr allgemeinen Darstellung ist es möglich, zu beweisen, daß ein Netzwerk, das neben RLCÜF-Netzwerkelementen aus passiven nichtlinearen zeitvarianten Widerständen besteht eindeutige Lösungen aufweist und mit unserem Verfahren analysiert werden kann, wenn die Netzwerkerregungen sinnvoll angebracht sind. Die Gleichstromanalyse elektrischer Netzwerke der genannten Klasse stellt einen trivialen Sonderfall unserer Theorie dar.

Den Abschluß dieser Arbeit bildet die Umsetzung des Iterationsverfahrens in ein numerisches Verfahren zur praktischen Berechnung der gesuchten Systemreaktion auf dem Digitalrechner, wobei wir uns auf die Untersuchung von Übergangsvorgängen beschränken. Anhand einer Fehleranalyse sind wir in der Lage, relativ leicht nachprüfbare Fehlerschranken für die unvermeidlichen Berechnungsfehler anzugeben. Die gute Anwendbarkeit sowohl unserer Theorie, als auch unseres numerischen Verfahrens wird mittels verschiedener praktischer Beispiele aus der Systemtheorie, Netzwerktheorie und Regelungstechnik demonstriert. Dabei besteht die Analyse des Systems nicht nur aus einer Berechnung der gesuchten Systemreaktion bei vorgegebener Systemerregung, sondern auch in der Lieferung allgemeiner Aussagen wie zum Beispiel die Stetigkeit des Rückkopplungssystems und eindeutige Lösbarkeit der Systemgleichung.

Für weitere Untersuchungen auf der Basis der hier beschriebenen Methoden gibt es mehrere Zielrichtungen:

Es liegt nahe die Impulsantwort des linearen Systems **H** derart zu modifizieren, daß sie nicht nur einen Deltastöß im Zeitnullpunkt sondern einen ganzen Impulskamm enthalten darf. Ferner darf vermutet werden, daß sich alle neuen Ansätze in dieser Arbeit auf die Stabilitätsuntersuchung von Systemen der hier beschriebenen Klasse übertragen lassen, da die Frage nach der Systemstabilität mit ganz ähnlich gearteten funktionalanalytischen Methoden behandelt werden kann. Abschließend sei erwähnt, daß die in dieser Arbeit angewandten Fixpunktverfahren auch für andere ingenieurwissenschaftliche Problemstellungen verwendet werden können (optimale Regelprozesse).

.

Ich danke Herrn Prof. Dr.-Ing. R. Unbehauen für die Themenstellung und Unterstützung dieser Arbeit.
Herrn Dr. rer. nat. R. Böhme danke ich für sein Interesse an dieser Arbeit und für die Übernahme des **Korreferats**.

LITERATURVERZEICHNIS

- [1] F.S. Acton: Numerical methods that work. Harper & Row, New York 1970.
- [2] B.D.O. Anderson, S. Vongpanitlerd: Network analysis and synthesis. Prentice-Hall, Englewood Cliffs 1973.
- [3] B. Gaißmaier: Schnelle und genaue Berechnung von Fourierintegralen durch Kombination von Spline-Interpolation und schneller Fourier-Transformation. NTZ 24(1971) S. 601 - 605.
- [4] J.M. Holtzman : Nonlinear system theory . Prentice-Hall, Englewood Cliffs 1970.
- [5] A.W. Naylor, G.R. Sell: Linear operator theory. Holt, Rinehart and Winston 1971.
- [6] R.W. Newcomb: Linear multiport synthesis. Mc Graw-Hill, New York 1966.
- [7] W.A. Porter: Modern foundations of system engineering. Macmillan Company, New York 1966.
- [8] F. Riesz, B.S. Nagy: Vorlesungen über Funktionalanalysis. VEB Deutscher Verlag der Wissenschaften, Berlin 1969.
- [9] I.W. Sandberg: On the properties of some systems that distort signals-I. B.S.T.J. 42, 1963, S. 2033 - 2047.
- [10] I.W. Sandberg: On the properties of some systems that distort signals-II. B.S.T.J. 43, 1964, S. 91 - 112.
- [11] I.W. Sandberg: Signal distortion in nonlinear feedback systems. B.S.T.J. 42, 1963, S. 2533 - 2550.
- [12] I.W. Sandberg: On truncation techniques in the approximate analysis of periodically time-varying nonlinear networks. Proceedings of the Allerton conference on circuit and system theory, University of Illinois, 1963.
- [13] I.W. Sandberg: On the response of nonlinear control systems to periodic input signals. B.S.T.J. 43, 1964, S. 911 - 926.
- [14] I.W. Sandberg: On truncation techniques in the approximate analysis of periodically time-varying nonlinear networks. IEEE Trans. CT-11, 1964, S.195 - 201.

- [15] I.W. Sandberg: A note on the application of the contraction-mapping fixed-point theorem to a class of nonlinear functional equations. SIAM 7, Nr.2, 1965, S. 199 - 204.
- [16] I.W. Sandberg: On the L_2 -boundedness of solutions of nonlinear functional equations. B.S.T.J. 43, 1964, S. 1581 - 1599.
- [17] I.W. Sandberg: On the boundedness of solutions of nonlinear integral equations. B.S.T.J. 44, 1965, S. 439-453.
- [18] I.W. Sandberg: A frequency-domain condition for the stability of feedback systems containing a single time-varying nonlinear element. B.S.T.J. 43, 1964, S. 1601 - 1608.
- [19] I.W. Sandberg: Some results on the theory of physical systems governed by nonlinear functional equations. B.S.T.J. 44, 1965, S. 871 - 898.
- [20] I.W. Sandberg: A stability criterion for linear networks containing time-varying capacitors. IEEE Trans. CT-12, 1965, S. 2 -11.
- [21] E.C. Titchmarsh: Introduction to the theory of fourier integrals. Clarendon Press, Oxford, 1967.
- [22] R. Unbehauen: Systemtheorie. R. Oldenbourg Verlag, München 1971.
- [23] R. Unbehauen: Elektrische Netzwerke. Springer Verlag, Berlin 1972.
- [24] N. Wiener: The fourier integral and certain of its applications. Dover Publications, New York.
- [25] G. Zames: Functional analysis applied to nonlinear feedback systems. IEEE Trans. CT-10, S. 392 - 404.

LEBENS LAUF

- 21.7.1946 geboren in Kaufbeuren im Allgäu als zweites Kind des Ingenieurs Kurt Huber und seiner Ehefrau Magdalena geb. Sailer
- 1952 - 1963 Schulzeit in Kaufbeuren, Abschluß mit der Mittleren Reife an der Staatlichen Realschule für Knaben
- 1963 - 1965 Praktikantentätigkeit bei mehreren Industriefirmen im Allgäuer Raum
- 1965 - 1968 Studium der Elektrotechnik am Rudolf-Diesel-Polytechnikum der Stadt Augsburg, Abschluß als Ing. (grad.) und Erwerb der eingeschränkten Hochschulreife
- 1968 - 1971 Studium der Elektrotechnik an der Technischen Universität München, Diplomarbeit bei Prof. Saal über numerische Aspekte bei der Filtersynthese
- 1972 - 1975 Verwalter der Dienstgeschäfte eines wissenschaftlichen Assistenten, später wissenschaftlicher Assistent am Institut für Allgemeine und Theoretische Elektrotechnik der Universität Erlangen-Nürnberg, Arbeitsgebiete in der nichtlinearen Systemtheorie und Netzwerktheorie

Copie - Center 2000
852 ERLANGEN • Obere Karlstr. 30